République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

جامعة بجاية
Tasdawit n Bgayet
Université de Béjaïa

Université A. Mira de Béjaïa
Faculté des Sciences Exactes
Département d'Informatique

Mémoire de Fin d'Etude

En vue de l'obtention du diplôme de Master professionnel en Informatique

Option : Administration et sécurité des réseaux

Thème

Conception et réalisation d'un système de prédiction du vainqueur de la coupe du monde

Réalisé par :

M^elle^ BESSAOUDI Sabrina
Mr OUCHENE Raouf

Soutenu le 14 septembre devant le jury composé de :

Examinateur 1 :	Pr AMROUN Kamal	Professeur Université de Béjaïa.
Examinatrice 2 :	Dr ALOUI Soraya	M.C.A Université de Béjaïa.
Encadrante :	Dr EL BOUHISSI Houda	M.C.A Université de Béjaïa.

Promotion 2021 - 2022

Dédicace

La page de la dédicace est généralement celle qui est écrite en dernier et lue en premier, aussi je doute que ce mémoire puisse faire exception à la règle.

Nous dédions ce mémoire avec un énorme plaisir, un cœur ouvert et une immense joie :

A mes parents,

Il est naturel que ma pensée la plus forte aille vers mon cher père, décédé cette année, qui m'a toujours pousser et motivé pour s'avancer avec mes études, il été toujours pour moi un exemple du père respectueux, honnête, de la personne méticuleuse. Grâce à lui j'ai appris le sens du travail et de la responsabilité. Aucune dédicace ne saurait exprimer l'amour l'estime et le respect que j'ai toujours eu pour lui.

Ce modeste travail est le fruit de tous les sacrifices qu'il a déployés pour mon éducation et ma formation. Je t'aime papa et j'implore le tout-puissant t'accueil en son vaste paradis.

Ce travail est dédié aussi à la meilleure femme au monde ma mère, à qui je dois la vie et une part essentielle de ma personnalité. Qu'elle sache que l'amour qu'elle me donne continue à m'animer et me permet d'envisager l'avenir comme un défi.

A mes chères sœurs et mes frères,

En témoignage de notre sincère reconnaissance pour les efforts qu'ils ont consenti pour l'accomplissement de ce projet. On leur dédie ce modeste travail en témoignage de notre grand amour et notre gratitude infinie.

A mon ami B.Walid

Pour son aide et son soutien moral son encouragement durant l'élaboration du travail de fin d'étude en le souhaitant un brillant avenir.

Je ne peux pas trouver les mots justes et sincères pour t'exprimer mon affection et mes pensées, pour moi un frère et un ami sur qui je peux compter. En témoignage de l'amitié qui nous unit et des souvenirs de tous les moments que nous avons passés ensemble, je tu dédie ce travail et je tu souhaites une vie pleine de santé et de bonheur.

A toutes nos Familles

A tous ceux dont l'oubli du nom n'est guère celui du cœur ...

Bessaoudi Sabrina

Dédicace

A mes chers parents,

Pour tous leurs sacrifices, leur amour, leur tendresse, leur soutien et leurs prières tout au long de mes études,

***A mes chères sœurs**,*

Pour leurs encouragements permanents, et leur soutien moral,

A mes chers frères,

Pour leur appui et leur encouragement,

A ma binôme et sa famille

***A toute ma famille**, pour leur soutien tout au long de mon parcours universitaire,*

Que ce travail soit l'accomplissement de vos vœux tant allégués, et le fuit de votre soutien infaillible,

Merci d'être toujours là pour moi.

Ochene Raouf

Remerciement

On dit souvent que le trajet est aussi important que la destination. Les cinq ans de master nous ont permis de bien comprendre la signification de cette phrase toute simple. Ce parcours, en effet, ne s'est pas réalisé sans défis et sans soulever de nombreuses questions pour lesquelles les réponses nécessitent de longues heures de travail.

Nous tenons à la fin de ce travail à remercier ALLAH le tout puissant de nous avoir donné la foi et de nous avoir permis d'en arriver là.

En premier lieu nous tenons à remercier profondément notre promoteur ***Dr. EL BOUHISSI Houda*** *d'avoir accepté de nous encadrer, Un très grand merci pour sa compréhension, sa disponibilité, sa gentillesse et son temps précieux qu'elle nous a accordé afin de pouvoir soutenir notre mémoire.*

Nous souhaitons également faire part de nos reconnaissances à tous les enseignants qui nous ont éclairés la voie du savoir durant notre cursus. Un grand merci également à l'ensemble du personnel pédagogique, technique et administratif du département.

Nous tenons également à exprimer nos remerciements aux membres du jury, qui ont accepté d'évaluer notre travail.

Nous remercions infiniment ***Pr. AMROUN Kamal*** *d'avoir accepté de présidence mémoire. Nous la remercions pour le temps qu'elle a consacré pour lire et évaluer notre travail et pour son déplacement afin de participer au jury.*

Nous remercions également ***Dr. ALOUI Soraya*** *d'avoir consacré du temps pour lire et examiner notre mémoire.*

Nous clôturons cette liste de remerciements avec une pensée très affectueuse pour nos familles, qui nous ont toujours soutenus. Merci à nos parents, nos frères, nos sœurs, pour leurs encouragements, même à distance, et leur motivation sans faille.

Nous exprimons notre profonde gratitude à tous celles et ceux qui nous ont apporté leurs soutiens, leur amitié ou leur expérience tout au long de ce travail.

Table des matières

Table des figures

Liste des abréviations

CSV	**C**omma-**S**eparated **V**alues
DT	**D**ecision **T**ree
ET	**E**xtra **T**ime
FN	**F**aux **N**égatif
FP	**F**aux **P**ositif
IA	**I**ntelligence **A**rtificielle
IDE	**I**ntegrated **D**evelopment **E**nvironment
KNN	**K**-**N**earst **N**eighbors
MCS	**M**onte **C**arlo **S**imulation
MIT	**M**assachusetts **I**nstitute of **T**echnology
MLA	**M**achine **L**earning **A**lgorithm
NB	**N**aïve **B**ayes
NN	**N**eural **N**etworks
NumPy	**N**umerical **P**ython
Pandas	**P**anel **D**ata
PIL	**P**ython **I**maging **L**ibrary
POO	**P**rogrammation **O**rientée **O**bjet
RBF	**R**adial **B**asis **F**unction
RF	**R**andom **F**orest
SML	**S**upervised **M**achine **L**earning
SVM	**S**upport **V**ector **M**achine

MC	**M**onte **C**arlo
Tkinter	**T**ool **k**it **I**nterface
UEFA	**U**nion of **E**uropean **F**ootball **A**ssociations
VN	**V**rai **N**égatif
VP	**V**rai **P**ositif

Liste des tableaux

Introduction générale

Introduction générale

Si vous ne revenez pas de Mars ou d'une très longue hibernation, alors vous aurez certainement déjà entendu son nom: Paul le poulpe, l'oracle du Mondial. Peut-être en passe de devenir le poulpe le plus célèbre de la terre et des mers.
Paul, le poulpe qui s'est fait une réputation mondiale en prédisant avec une grande précision les résultats du Mondial de football.

Celui qui, pendant toute la compétition, a prédit sans jamais se tromper le résultat de chacun des matchs de l'équipe d'Allemagne lors de la coupe du monde de football en 2010, son mode d'opération pour partager ses prédictions avec les humains est de choisir une des différentes deux récipients transparents percés de trous, munis d'un couvercle, contenant chacun une moule décoquillée qu'on lui présente, chacun portant un drapeau d'une équipe.
Ce poulpe envoûta les Internautes en prédisant simplement les résultats de la Coupe du monde.

Vu qu'il est aimable de savoir les choses avant qu'elles arrivent, le poulpe Paul nous a motivé de créer un système de prédiction du vainqueur de la coupe de monde pour l'année 2022 ainsi que les autres coupes du monde et de prévision du résultat des matches en introduisant simplement les équipes en compétition .

Vu que le football est le sport le plus médiatisé de la planète, l'engouement qu'il provoque soulève des stades entiers mais également des sommes impressionnantes. En effet, de nombreux business fructueux tournent autour du ballon rond, et l'issue d'un seul match peut avoir des retombées financières importantes. Les jeux d'argent sont un des aspects les plus lucratifs liés à ce sport, avec le principe des paris sportifs. Lors de la coupe du monde de foot en 2018, plus de 690 millions d'euros ont été pariés par les particuliers en France (source Arjel), la question qui nous vient à l'esprit peut-on vraiment prédire le vainqueur de la coupe du monde [1] ?

Nous sommes bien conscients que nous n'arriverons pas à prédire avec une précision infaillible l'issue de tous les matchs, car ces derniers ont une part de surprise et d'imprévisible conséquente. Nous espérons tout de même obtenir des résultats de prédictions du niveau d'un être humain qui suit le Football régulièrement et connaît bien la plupart des équipes [1].

Si certains d'entre vous ont l'âme d'un parieur et souhaitent éclairer leurs décisions par des statistiques, Nous avons créé un modèle de prédiction qui permet de générer les côtes et prédire l'équipe gagnante d'un match de la coupe du monde, notre modèle auras donc pour but de prédire l'issue des matchs en récoltent des données viennent principalement de site « www.fifa.com », sachant que nous aurons ainsi besoin d'un maximum de données sur les matchs et les équipes et que nous avons besoin de données les plus complètes possibles. Par chance, le football étant le sport le plus suivi de la planète, on peut trouver des jeux de données d'assez bonnes qualités, puis évaluer les données collectées afin qu'il soit possible de créer un modèle d'apprentissage automatique supervisé pour prédire le résultat des matchs de la coupe du monde.

En pratique, notre système est conçu pour l'analyse et la prédiction d'un championnat de football. Il est basé sur la loi de Poisson et inclue aussi les points des équipes comme covariables et incorporent les différences d'effets spécifiques à l'équipe. Ce modèle de prédiction de la Coupe du Monde de la FIFA 2022 et les futurs coupes du monde est ajusté sur tous les matchs de football sur terrain neutre des équipes participantes depuis l'apparition de la coupe du monde. Sur la base de ce modèle pour les matchs simples, nous utilisons des simulations de Monte-Carlo pour estimer les probabilités d'atteindre les différentes étapes de la Coupe du monde de football 2022 pour toutes les équipes sachant que le modèle favorise que le Brésil est le nouveau champion du monde de la FIFA , ainsi que notre système prédire l'issue de chaque match de la coupe du monde .

Pour toutes ces raisons l'objectif principal de ce travail est la réalisation d'un système de prédiction de vainqueur de la coupe du monde.

Pour atteindre notre objectif, nous dresserons un plan simple et cohérant à nos yeux ; composé de trois chapitres :

Le premier chapitre, est consacré aux généralités sur la prédiction, systèmes de prédiction existants, rôle, importance, fonctionnement, les avantages et les inconvénients, ainsi que la prédiction pour la coupe du monde et on se termine par les problèmes et les intérêts.

Le deuxième chapitre servira à la conception et la réalisation de notre système de prédiction où nous allons présenter le schéma qui décrit notre algorithme de classification

« algorithme de monte Carlo », ainsi que des quelques captures d'écrans pour notre algorithme.

Dans le troisième chapitre, nous présentons un état de l'art sur les méthodes prédictives d'apprentissage automatique supervisé et faire une étude comparative entre eux.

Dans le quatrième, chapitre, nous présenterons d'abord les outils et les langages de programmation utilisés ainsi quelques librairies. Nous finalisons notre travail avec la présentation de notre application où nous allons présenter leur fonctionnement de notre à travers quelques interfaces.

Nous terminons ce manuscrit par une conclusion sur nos travaux ainsi que des perspectives de recherche à court et long terme.

Chapitre 1 : Introduction aux prédictions

Chapitre1 : Introduction aux prédictions.

1.1. Introduction

Les prédictions ont aujourd'hui une place majeure dans plusieurs domaines tel que les paris sportifs notamment le football, Elles sont utilisées pour prédire l'issue des matchs de la coupe du monde.

En bref, la prédiction est une technique statistique utilisant l'apprentissage automatique et l'exploration de données pour prédire et prévoir les résultats futurs probables à l'aide de données historiques et existantes. Elle fonctionne en analysant les données actuelles et historiques et en projetant ce qu'il apprend sur un modèle généré pour prévoir les résultats probables [8].
Dans ce chapitre, nous allons présenter les différentes notions liées aux prédictions.

- Dans la première partie : nous allons fait un tour d'horizon sur la prédiction mettant l'accent sur les systèmes de prédiction leur rôle, importance, fonctionnement, leur avantages et inconvénients.
- Dans la deuxième partie : nous abordons la prédiction pour la coupe du monde, ses problèmes et ses intérêts.

Enfin nous terminons par une conclusion.

1.2. Généralités sur la prédiction

1.2.1. Apparition des prédictions et ses notions

C'est un fait, l'homme cherche à explorer l'avenir. Nous sommes passés du temps des prophéties au temps des sciences. La prédiction n'est pas une science exacte, ce qui n'empêche pas les chercheurs d'avoir posé des fondements pour avancer sur le sujet. Attardons nous sur la vision qu'en ont les économistes et les mathématiciens.

Avant d'introduire ce qu'est un système de prédiction, définissons ce que nous entendons par prédiction. La prédiction repose sur l'analyse prédictive qui est une branche de machine Learning. Pour ce faire, on utilise des algorithmes et modèles statistiques qui analysent les données historiques pour y découvrir des tendances, qui seront ensuite extrapolées pour faire de la prédiction [2].

La prédiction ou l'analyse prédictive révolutionnera la recherche et le développement dans n'importe quel domaine donné. La prédiction existe depuis plus de 75 ans, mais vient tout juste d'atteindre le statut de grand public. Il est actuellement utilisé dans tous les secteurs et domaines fonctionnels, elle a été apparue lorsque les gouvernements ont commencé à utiliser les premiers modèles informatiques. Avec la programmation non linéaire et l'analyse en temps réel, l'analyse de données et l'analyse prescriptive se généralisent et deviennent disponibles pour toutes les organisations. Avec l'essor des technologies de Big data, nous sommes maintenant entrés dans une nouvelle ère d'analyse prédictive qui personnalisera et démocratisera les données (analytiques) pour les organisations, les individus et les gouvernements [8].

Selon la littérature, l'histoire des prédictions passe par plusieurs périodes :

L'analyse prédictive trouve ses origines en 1940, lorsque les gouvernements ont commencé à utiliser les premiers modèles informatiques par exemple les simulations de Monte-Carlo… elle est utilisée aussi pendant la deuxième guerre mondiale pour décoder les messages allemands à automatiser le ciblage des armes antiaériennes utilisées contre les avions ennemis et à recourir aux simulations informatiques pour prédire le comportement des réactions nucléaires en chaîne du projet Manhattan. Dans les années 1960, avec la programmation non linéaire et l'analyse en temps réel, l'analyse de données et l'analyse prescriptive se généralisent et deviennent disponibles pour toutes les organisations. Puis dans les années 1970– 1990, l'analyse a été utilisée de façon plus répandue dans les entreprises et l'analyse prescriptive en temps réel est devenue réalité dans les startups technologiques. Toutefois, l'analyse prédictive est largement restée l'apanage des statisticiens. Aujourd'hui, l'analyse prédictive a finalement investi la majorité des entreprises, et avec l'essor des technologies de big data, nous sommes maintenant entrés dans une nouvelle ère d'analyse prédictive qui personnalisera et démocratisera les données (analytiques) pour les organisations, les individus et les gouvernements [10,11].

1.2.2. Définition de la prédiction

Malgré l'existence de plusieurs définitions pour les prédictions, celle qui reste référence et certainement la plus populaire, parmi celles qui existent dans la littérature est que la prédiction est la technique dans l'analyse sera analytique et statistique. La prédiction ressemble à la classification et à l'estimation mais dans une échelle temporelle différente. Elle

s'appuie à la fois sur des données actuelles et historiques, permet de créer des hypothèses et des prédictions sur des événements futurs. La seule méthode pour mesurer la qualité de la prédiction est d'attendre [2].

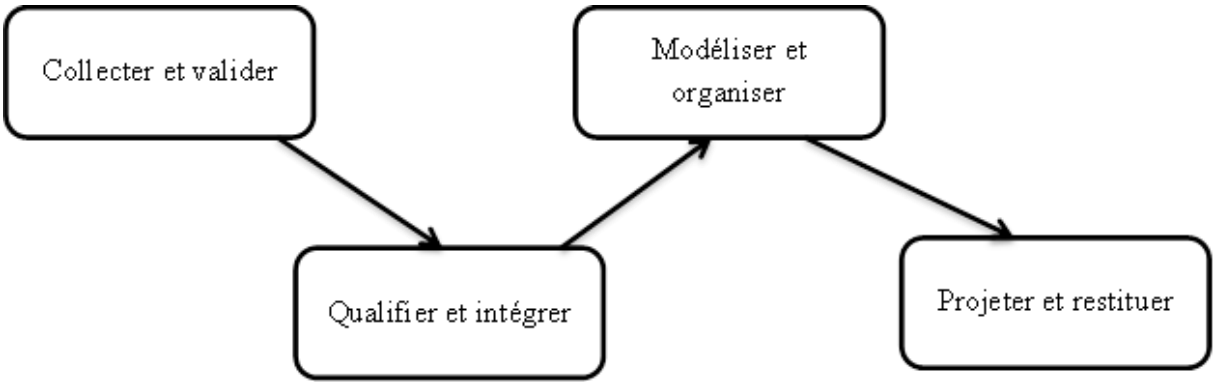


Figure I. 1 Le processus de prédiction [4].

1.2.3. Les méthodes de prédiction

On entend par méthodes de prédiction, l'ensemble des techniques, outils et méthodologies permettant d'établir des prédictions. Certaines méthodes les plus populaires sont les suivantes:

1.2.3.1. Arbres de décision

Les arbres de décision sont un modèle de classification/prédiction simple d'utilisation qui permet une interprétation très utile aux experts.

Un arbre de décision est formellement une structure d'arbre comprenant des nœuds, des branches et des feuilles (ou nœuds terminaux).

Les algorithmes d'arbre de décision représentent graphiquement des données (exploitées, open source, internes) en branches pour afficher les résultats possibles de diverses décisions.et ils classent les variables de réponse et prédisent les variables de réponse en fonction des décisions passées, peuvent être utilisés avec des ensembles de données incomplets et sont facilement explicables et accessibles pour les scientifiques de données novices [9].

1.2.3.2. Analyse des séries chronologiques

Il s'agit d'une technique de prédiction d'événements à travers une séquence de temps. Autrement dit prédire les événements futurs en analysant les tendances passées et en extrapolant à partir de là.

Se définit aussi comme un ensemble de données ou d'observations qui se réfèrent à une ou plusieurs variables et sont classées chronologiquement.

Les séries chronologiques sont très importantes en économie. Car, en économie, presque toutes les variables sont collectées dans le temps. En d'autres termes, il est intéressant de voir l'évolution d'une variable dans le temps, pas la valeur spécifique à un moment donné. Ainsi, chaque fois que des variables économiques sont analysées, on parle de cycles ou de tendances économiques.

L'ordre des données étant d'une importance vitale, il faut tenir compte du fait que cela modifie l'analyse et l'interprétation des données. L'économétrie, chargée de rechercher et d'estimer les relations entre les variables économiques, doit donc tenir compte de ce fait [3].

1.2.3.3. Régression logistique

Est une méthode d'analyse statistique qui aide à la préparation des données. On lui introduisant les données, l'algorithme les trier et à les classer ce qui améliore la capacité de l'algorithme et par conséquent, des prédictions peuvent être faites

La régression logistique est devenue un outil important dans la discipline de l'apprentissage automatique. Cette approche permet d'utiliser un algorithme dans l'application d'apprentissage automatique pour classer les données entrantes en fonction des données historiques. Plus il y a de données pertinentes en entrée, plus l'algorithme est en mesure de prédire des classifications au sein des jeux de données.

La régression logistique peut également jouer un rôle dans la préparation des données, en permettant aux jeux de données d'être répartis dans des catégories spécialement définies au cours du processus d'extraction, transformation et chargement (ETL, Extract, Transform, Load) afin d'organiser les informations aux fins d'analyse [3].

1.2.3.4. Les réseaux de neurones

Les réseaux de neurones sont des modèles d'apprentissage automatique capables de représenter une relation entre des données d'un espace *X* et un espace de sortie *Y*. C'est une méthode examine de grands volumes de données étiquetées à la recherche de corrélations entre les variables dans les données. Les réseaux de neurones constituent la base de nombreux

exemples actuels d'intelligence artificielle (IA), notamment la reconnaissance d'images , les assistants intelligents et la génération de langage naturel.

L'unité de calcul de base est le neurone. Celui-ci prend en entrée plusieurs signaux et les interprète pour envoyer un nouveau signal vers d'autres neurones ou vers la sortie du réseau de neurones, c'est-à-dire la sortie du modèle [3].

Grâce aux réseaux de neurones nous pouvons résoudre plusieurs problèmes tels que [3] :

- Classification.
- Catégorisation.
- Approximation de fonction.
- Prédiction/ prévision.
- Optimisation.
- Contrôle.

1.2.4. Les algorithmes de prédiction

Les algorithmes prédictifs sont des modèles mathématiques combiné avec une grande quantité de données pour anticiper les résultats futurs les plus sûrs en tenant compte des résultats survenus[14].Certains algorithmes les plus populaires sont les suivantes :

1.2.4.1. Random Forest ou les forêts aléatoires

Proposées par Breiman en 2001 forme une famille de méthodes de classification qui se base sur l'assemblage de plusieurs arbres de décision. Il utilise la classification et la régression pour organiser et étiqueter de grandes quantités de données.

La prédiction dans cet algorithme est faite en combinant les prédictions de chaque arbre (par *vote* par exemple). Les conditions associées aux nœuds peuvent concerner plusieurs attributs (tirés aléatoirement) combinés éventuellement de façon non linéaire.

L'efficacité prédictive des *forêts aléatoires* et leurs propriétés mathématiques en font des classifieurs très puissants et donc très populaires, notamment en apprentissage automatique [9].

1.2.4.2. Gradient boosted model ou modèle boosté en dégradé

Semblable à Random Forest, c'est un modèle dit ensemblistes, qui reposent sur une approche séquentielle dite adaptative qui consiste à combiner plusieurs modèles avec des pouvoirs de prédiction faibles pour obtenir un modèle avec un pouvoir de prédiction puissant.

Aussi le boosting de gradient est une méthode utilisée dans le machine learning pour réduire les erreurs dans l'analyse prédictive de données. Les scientifiques des données utilisent des données étiquetées pour entraîner des logiciels de machine learning (appelés modèles de machine learning) à faire des prédiction sur des données non étiquetées. Un modèle de machine learning unique peut faire des erreurs de prédiction selon la précision du jeu de données d'entraînement. Par exemple, si un modèle d'identification de chats n'a été entraîné que sur des images de chats blancs, il peut occasionnellement faire des erreurs lors de l'identification d'un chat noir. Le boosting s'efforce de résoudre ce problème en entraînant successivement plusieurs modèles afin d'améliorer la précision du système global [3].

1.2.4.3. K-Means

Cet algorithme regroupe les points de données de la même manière que les modèles de clustering et est populaire dans la conception d'offres de vente au détail personnalisées. Il crée des offres personnalisées en recherchant des similitudes entre de grands groupes de clients.

C'est l'un des algorithmes de clustering les plus répandus. Il permet d'analyser un jeu de données *caractérisées* par un ensemble de descripteurs, afin de regrouper les données "similaires" en groupes (ou clusters) [3].

1.2.4.4. Prophète

Procédure de prévision, cet algorithme est particulièrement efficace lorsqu'il s'agit de planifier la capacité. Cet algorithme traite des données de séries chronologiques et est relativement flexible.

Aussi le prophète est un modèle de régression supplémentaire sous forme d'une courbe tendancielle de croissance logistique ou linéaire. Il inclut un composant saisonnier annuel modélisé sur les séries de Fourier et un composant saisonnier hebdomadaire modélisé modélisée à l'aide de variables fictives [3].

1.2.5. Les langages et les logiciels pour servir les prédictions

On présentera donc, dans une première partie, les langages les plus utilisés pour la programmation des algorithmes prédictifs de machine Learning. Dans la deuxième partie, nous avons proposé quelques exemples de logiciels en lien direct avec ces langages.

1.2.5.1. Présentation de langage python

Python est un langage de programmation adapté aux analyses statistiques et au Machine Learning, aussi il est plus généraliste on pourra l'utiliser également pour d'autres tâches informatiques. Est un langage actuellement plus actif en termes de développement [13].

Figure I. 2 Logo du langage Python.

1.2.5.2. Présentation de langage R

R est un langage de programmation adapté aux analyses statistiques et au Machine Learning, il existe depuis 1993 et dont le développement a été fortement accéléré dès les années 2000.Est un langage disposant de plus d'algorithmes notamment sur les aspects de série temporelle [12].

Figure I. 3 Logo du logiciel R.

1.2.5.3. Les logiciels en lien direct avec Le R et le python

Présentation de Spyder

Spyder **(**nommé Pydee dans ses premières versions) est un environnement de développement pour Python. Libre(Licence MIT) et multiplateforme (Windows, Mac OS, GNU/Linux),il intègre de nombreuses bibliothèques d'usage scientifique:Matplotlib,NumPy,SciPy et IPython [13] .

Figure I. 4 Logo d'environnement Spyder.

Présentation de RStudio

RStudio est un environnement de développement intégré qui permet de travailler en R, développer de nouvelles bibliothèques et travailler avec des notebooks [12].

Figure I. 5 Logo d'environnement RStudio.

1.3. Systèmes de prédiction

Les systèmes prédictifs n'ont pas besoin d'être créés à partir de zéro pour chaque application. Les outils d'analyse prédictive utilisent une variété de modèles et d'algorithmes approuvés qui peuvent être appliqués à un large éventail de cas d'utilisation ; autrement dit un système de prédiction est basé sur des algorithmes et des données existants. Parmi les principaux systèmes de prédiction On trouve :

1.3.1. Système de classification

Considéré comme le système le plus simple, il catégorise les données pour une réponse simple et directe aux requêtes [3].

1.3.2. Système de clustering

Ce système imbrique les données par des attributs communs. Il fonctionne en regroupant des choses ou des personnes ayant des caractéristiques ou des comportements communs et planifie des stratégies pour chaque groupe à une plus grande échelle [3].

1.3.3. Système de prévision

Il s'agit d'un système très populaire, et il fonctionne sur tout ce qui a une valeur numérique basée sur l'apprentissage à partir de données historiques [3].

1.3.4. Système des valeurs aberrantes

Ce système fonctionne en analysant des points de données anormaux ou aberrants [3].

1.3.5. Système de série chronologique

Ce système évalue une séquence de points de données en fonction du temps [3].

1.4. Le rôle et l'importance

- Un système de prédiction est conçu dans le but d'anticiper les comportements et les actions futures ; à côté d'autres bénéfices tels que :
- Prédiction des comportements.
- Réduit le temps, les efforts et les coûts de prévision des résultats.
- **Diminution des risques :** Les données intégrées à un modèle prédictif facilitent la prise de décision en permettant d'évaluer les conséquences d'un choix. Les risques de mauvaise décision sont alors fortement limités [5].

1.5. Fonctionnement de la prédiction

En global, notre algorithme prédictif se base sur les algorithmes de machine Learning. Premièrement, On donne à l'algorithme des données d'entrainement. Puis, l'algorithme d'apprentissage machine apprend un modèle capable de généraliser à de nouvelles données.

Le principe de notre algorithme prédictif se base essentiellement sur :

1) Acquisition et stockage des données

Afin de réaliser nos prédictions, nous avons besoin de données les plus complètes possibles. C'est le cas du site www.fifa.com , sachant que plus un modèle dispose de données d'entraînement, plus il peut correctement prédire des observations [1].

2) Prétraitement de données

On transforme les données de sorte qu'elles soient utilisables par notre algorithme de prédiction. (Transformer les fichiers Csv en fichiers Excel comme exemple) [1].

3) Choix d'une méthode de prédiction

Une fois nos données à un format utilisable, nous avons essayé d'appliquer nos premiers modèles de machine Learning.

Nous avons d'abord dû faire le choix des modèles à appliquer. Compte tenu de notre quantité de données assez limitée (même si l'on s'entraîne sur 2 ou 3 ans on dépasse à peine le millier de matchs), un réseau de neurones n'était clairement pas envisageable. Nous avons donc plutôt considéré des classifier fonctionnant bien avec des quantités de données plus faibles : SVM, Régression Logistique, Arbre de décision et Random Forest [1].

4) Développement de la méthode choisie [1].

1.6. Points forts

- Efficace, c'est devenu une bonne affaire en raison du nombre croissant de parieurs et de passionnés de sport qui ont besoin de conseils de paris et de pronostics.
- Gagner de l'argent : la façon de vivre de certains qui ont eu la chance de gagner assez d'argent.
- Ces systèmes se caractérisent aussi par leur facilité et rapidité. En plus, ils ne coûtent pas cher et ils sont très faciles à réaliser.
- Puissant et précis, bonnes performances sur de nombreux problèmes.
- Disponibilité des données.
- N'implique pas les rigueurs de la sélection et du nettoyage des données [7].

1.7. Points faibles

- Ces systèmes sont probabilistes : ils n'estiment pas l'issue d'un domaine donné avec une précision infaillible.
- Complexe et pouvant entraîner des erreurs.
- Difficile d'identifier les erreurs.
- Ils ont une faible prédiction.
- Ils prennent du temps lors de leur exécution.
- Capacités limitées.

- Tâche fastidieuse de nettoyage des données.
- Sont Coûteux.
- Pas le meilleur choix pour un grand nombre de données [7].

1.8. Prédiction pour la coupe du monde

La génération de prédictions pour les scores de football est un thème de recherche important depuis le milieu du 20e siècle, avec les premières approches de modélisation statistique et des idées provenant de Moroney (1956) et Reep (1971) [7].

La prédiction statistique des résultats de football pour la coupe du monde est une méthode utilisée pour les paris sportifs afin de prédire l'issue des matchs de la coupe du monde à l'aide d'outils statistiques.

Les prédictions utilisent diverses méthodes, telles que l'analyse des cotes des bookmakers et l'utilisation de résultats et de classements récents. Les prédictions peuvent également être un modèle basé sur les résultats qui modélise directement la probabilité d'un résultat de jeu (victoire, défaite ou match nul), par rapport à un modèle basé sur le score, qui se concentre sur le score du match. Le modèle de score de match convient mieux à un tournoi tel que la Coupe du Monde de la FIFA, car les scores de match sont importants en phase de groupes pour déterminer quelles équipes progressent. Beaucoup d'entre eux utilisent des simulations de Monte Carlo, ce qui signifie qu'ils simulent le tournoi des milliers de fois, et la probabilité qu'une équipe gagne le tournoi représente la part des simulations dans lesquelles elle le remporte [5].

1.8.1. Problèmes

Malgré leur viralité, ces prédictions n’engagent que leur auteur puisque personne ne sait encore à ce jour quelles équipes vont se qualifier après les barrages. Et personne ne peut prédire les classements, car tout cela se fait au hasard. Il est encore, statistiquement, très difficile de prédire les scores de plusieurs matches jusqu’à la finale. Même les plus puissants des calculateurs ne peuvent y arriver [6].

1.8.2. Intérêts

- Réduire l’incertitude liée au non connaissance du futur, en plus de susciter l'enthousiasme national.

- Pour des intérêts financiers : à l'issue d'un seul match peut avoir des retombées financières importantes.
- Il permet aux personnes d'acquérir des connaissances cachées et de les rendre plus compétitives.
- Grand intérêt pour le football dans le monde, en particulier les grands tournois internationaux tels que la Coupe du monde de la FIFA.
- également l'intérêt pour les paris liés au football et la prédiction de match/ligue

1.9. Conclusion

La prédiction est une approche présente dans plusieurs domaines quel que soit la médecine, les grandes entreprises, la météo ainsi que le football. Elle est essentiellement utilisée pour certains objectifs chacun au détriment de domaine dans laquelle est utilisée. Dans le côté programmation elle est basée sur plusieurs langages tels que le R et Python ainsi que Le spyder et le jupyter notebook comme logiciels en lien directe avec les langages de programmation mentionnés. Et malgré les méthodes et les langages performants utilisés pour la prédiction, nous sommes bien conscients que nous n'arriverons pas à prédire avec une précision infaillible l'issue de tous les matchs, car ces derniers ont une part de surprise et d'imprévisible conséquente.

Dans ce chapitre, nous avons abordé les différentes notions basiques qui sont très nécessaires pour voir c'est quoi une prédiction et ses différents systèmes et leurs fonctionnements. Puis, on traite les différents avantages et inconvénients de cette méthode ainsi que la prédiction pour la coupe du monde, ses intérêts et les différents problèmes rencontrés de la prédiction

La prédiction des matchs de football est devenue un domaine passionnant pour de nombreuses personnes dont le football est un domaine prolifique. .Pour cela on veut dans le prochain chapitre poursuivre la tendance de prédire les vainqueurs des matchs de la coupe du monde de la FIFA.

Chapitre 2 : Méthodes de prédiction

Chapitre 2 : Méthodes de prédiction.

2.1. Introduction

L'étude du corpus général et l'élaboration d'un mémoire de recherche nécessite généralement de mettre en exergue les différents concepts divergents du sujet de celui-ci, à fin de bien cerner nos connaissances. Pour cela dans ce chapitre nous allons présenter les différentes méthodes de la prédiction et puis on fera une étude comparative entre eux, en se faisant référence à une recherche bibliographique basées sur l'exploitation des documents (thèses, mémoires, articles scientifiques, livres…).

Parmi les différentes méthodes de prédiction on trouve les algorithmes d'apprentissage automatique qui utilisent une variété de méthodes statistiques, probabilistes et d'optimisation pour apprendre de l'expérience passée et détecter des modèles utiles à partir d'ensembles de données volumineux, non structurés et complexes [15].

Différents types d'algorithmes probabilistes jouent un rôle de plus en plus important dans l'informatique, en particulier dans la prédiction. A cet égard, les algorithmes d'apprentissage automatique supervisé ont été une méthode dominante dans le domaine de l'exploration de données, la prédiction a récemment montré un domaine d'application potentiel pour ces méthodes.

2.2. Algorithme d'apprentissage automatique supervisé pour la prédiction

L'apprentissage automatique a pour but de développer, d'analyser et d'implémenter des méthodes qui permettent à une machine d'évoluer grâce à un processus d'apprentissage, et de remplir des tâches qu'il est difficile ou impossible de remplir par des moyens algorithmiques plus classiques. Et avec l'alimentation de nouvelles données, les algorithmes d'apprentissage automatique ont tendance à faire des prédictions plus précises. (Figure 1) [16].

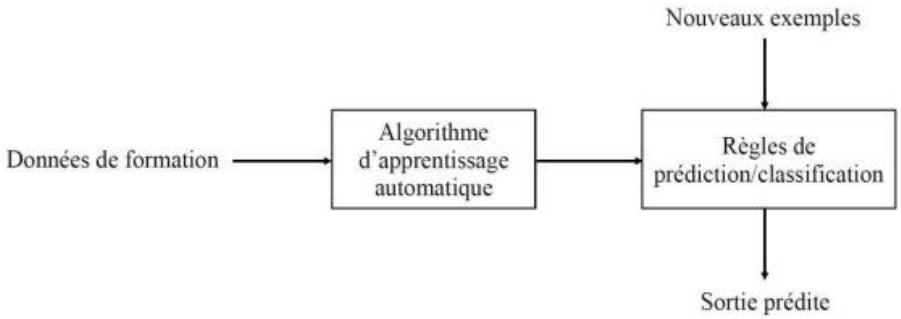


Figure II. 1 Approche typique de l'apprentissage automatique tirée de Liakos et al. (2018) [46].

Trois grandes approches relèvent de l'apprentissage automatique : supervisé, non supervisé et semi-supervisé [16].

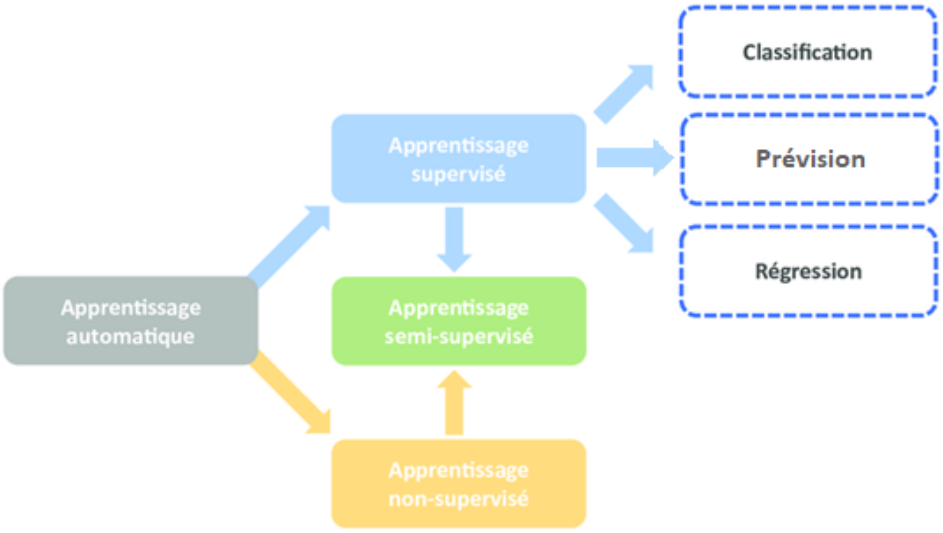


Figure II. 2 Les différents types d'algorithmes apprentissage automatique [16,17].

Dans les algorithmes d'apprentissage automatique supervisé (SML), un ensemble de données de formation étiqueté est d'abord utilisé pour former l'algorithme sous-jacent, autrement dit le SML est la recherche d'algorithmes qui raisonnent à partir d'instances fournies de manière externe pour produire des hypothèses générales, qui font ensuite des prédictions sur les instances futures. En comparaison avec les autres méthodes d'apprentissage automatique l'apprentissage automatique supervisé nécessite moins de données et facilite l'apprentissage car les résultats du modèle peuvent être comparés aux résultats réels marqués.

Ainsi que les algorithmes d'apprentissage supervisé conviennent bien à trois types: classification, régression et de prévision.

La classification : le programme d'apprentissage automatique lors classification doit tirer une conclusion à partir des valeurs observées et déterminer à quelle catégorie appartiennent les nouvelles observations. Par exemple, lors du filtrage des e-mails comme "spam" ou "non spam", le programme doit examiner les données d'observation existantes et filtrer les e-mails en conséquence

Régression : le programme d'apprentissage automatique lors de la régression doit estimer - et comprendre- les relations entre les variables. L'analyse de régression se concentre sur une variable dépendante et une série d'autres variables changeantes, ce qui la rend particulièrement utile pour la prédiction et la prévision.

Prévision : La prévision est le processus qui est utilisé pour faire des prédictions sur l'avenir sur la base des données passées et présentes, et est couramment utilisée pour analyser les tendances [17 ,48].

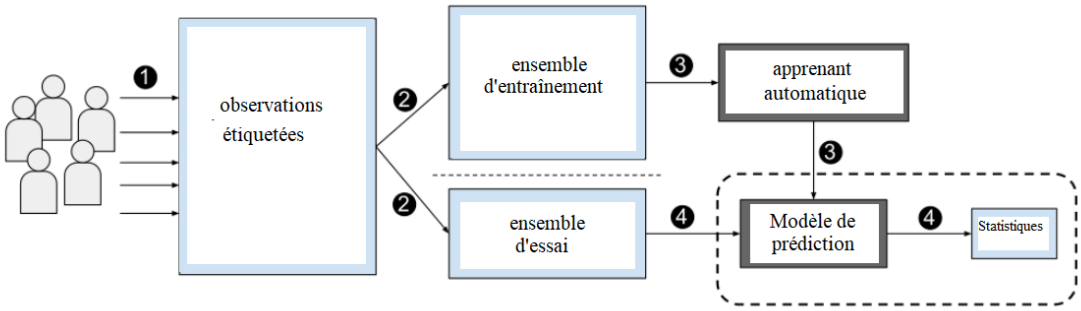


Figure II. 3 flux de travail d'un apprentissage supervisé pour la prédiction [53].

Les données d'un algorithme supervisé consistent en un ensemble de couples entrées / sorties {(X, Y)}. L'algorithme est formé en mappant les entrées sur les sorties (Y = f (X)). Lorsque vous fournissez une nouvelle entrée, l'algorithme devrait prédire la sortie. Autrement dit, étant donné un ensemble de caractéristiques {(x (1), …, x(m)} associés un ensemble de sorties {(y (1), …, y(m)} on veut construire un classifieur qui apprend à prédire y depuis x [18].

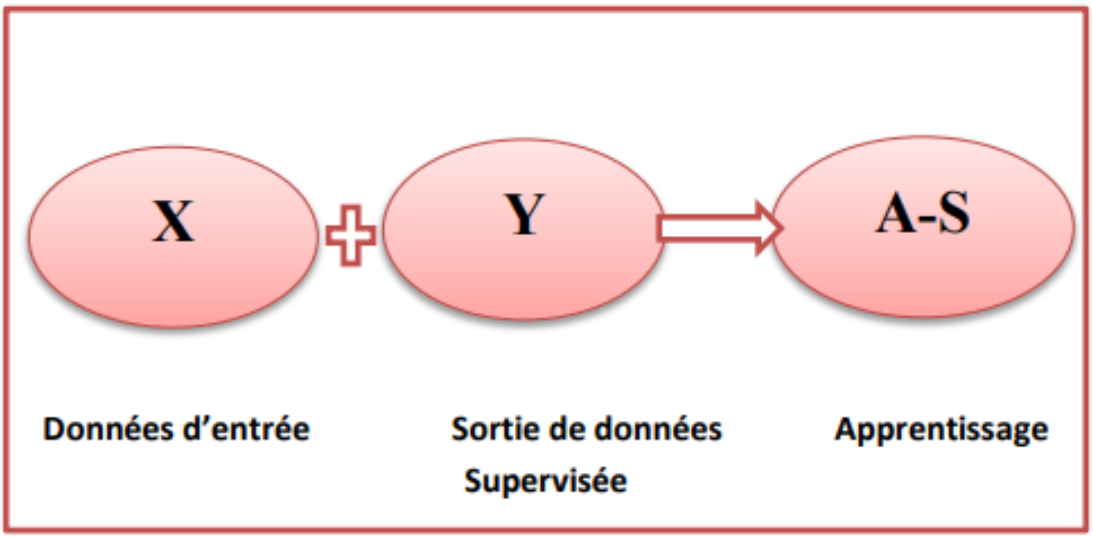


Figure II. 4 Apprentissage supervisé [48].

Parmi les algorithmes d'apprentissage supervisé, on peut citer machine à support de vecteur, k plus proches voisins, arbre de décision, naïves Bayes, random Forest.

2.2.1. Bayésien naïf NB (Algorithmes probabilistes)

NB est un algorithme assez intuitif à comprendre. Il se base sur le théorème de Bayes des probabilités conditionnelles suivant :

$P(H|E) = \frac{P(E|H) \cdot P(H)}{P(E)}$ Avec H une hypothèse, E une évidence et P(H) la probabilité à priori

Cette équation permet la prédiction sur la base de calcul probabiliste comparé avec plusieurs attributs afin de déterminer la valeur de la classe étudiée.

Naïve Bayes assume une hypothèse forte (naïve). En effet, il suppose que les variables sont indépendantes entre elles. Cela permet de simplifier le calcul des probabilités [18, 47]. A titre d'exemple l'algorithme de monte Carlo.

Il se résume suivant cet algorithme [47] :

Algorithme générique « naïve bayes »

Entrées :

Données d'entraînement T,

F = (f_1,f_2,$f_{3,..,}f_n$) // valeur de la variable prédictive dans l'ensemble de données de test.

Sorties :

Une classe d'ensemble de données de test.

Itérations :

1. Lire les données d'entraînement T ;
2. Calculer la moyenne et l'écart type des variables prédictives dans chaque classe ;
3. Répéter
   Calculer la probabilité de f_i en utilisant l'équation de densité de Gauss dans chaque classe ;
   Jusqu'à ce que la probabilité de toutes les variables prédictives (f_1,f_2,$f_{3,..,}f_n$) est calculée.
4. Calculer la vraisemblance pour chaque classe ;
5. Obtenir la plus grande probabilité ;

2.2.1.1. Avantages de méthode NB

1. Est une méthode rapide

2. facile et simple à implémenter [48].

3. ainsi qu'elle donne de bons résultats.

2.2.1.2. Inconvénients de méthode NB

Ses limites résident à l'hypothèse de fonctionnalités indépendantes. Dans la vraie vie, il est presque impossible d'obtenir un ensemble de fonctionnalités complétement indépendants [48].

2.2.2. Machines à vecteurs de support (SVM)

Est une méthode utilisée pour la classification et l'analyse de régression. En outre le SVM développé par Cortes et Vapnik, est un modèle d'apprentissage supervisé avec des algorithmes d'apprentissage associés qui analysent les données [20].

L'algorithme SVM est souvent utilisé pour l'obtention de meilleurs résultats que les autres classificateurs [21], ainsi que le problème pourrait ne pas être linéairement séparable [22]. Dans ce cas, un SVM avec un noyau non linéaire tel que la fonction de base radiale (RBF) serait approprié. Les SVM est utilisé dans le cas si l'on se trouve dans un espace de grande dimension [23].Ainsi que, les SVM est une extension du classificateur de vecteurs de support et est obtenu à la suite de l'élargissement de l'espace des caractéristiques d'une manière spécifique, à l'aide de noyaux [24].

La représentation du classificateur de vecteur de support linéaire est comme indiqué dans l'équation (1) [19] :

$$f(x) = \beta_0 + \sum_{i=1}^{n}(a_i\langle x, xi\rangle) \qquad (1)$$

Où $\alpha_1, \cdots, \alpha_n$ and β_0 sont des paramètres estimés par $\binom{n}{2}$ produits intérieurs $\langle xi, x'i\rangle$ entre toutes les paires d'observations d'entraînement. Remplacement du produit intérieur par $K(x_i,x'_i)$, où K est une fonction appelée le noyau. Le noyau linéaire est représenté comme indiqué dans l'équation (2) [19] :

$$K(x_i, x'_i) = \sum_{j=1}^{p}(x_{ij}, x'_{ij}) \qquad (2)$$

Le noyau polynomial de degré d (où d est positif) peut être représenté comme indiqué dans l'équation (3) [19] :

$$k(x_i, x'_i) = \left(1 + \sum_{j=1}^{p} x_{ij}x'_{ij}\right)^d \qquad (3)$$

Les résultats de la classification de la combinaison du noyau non linéaire et du classificateur de vecteur de support sont appelés SVM (équation (3)).

Le SVM est un algorithme d'apprentissage supervisé qui classe les données en deux classes ou plus en se basant sur le noyau et lorsqu'il y a un grand nombre d'exemples d'apprentissage, ce genre de classificateur n'est pas recommandé, est utilisé pour la classification binaire [20].

Les performances du classificateur SVM reposent sur le choix de contrainte de boîte et du paramètre du noyau, autrement dit paramètre de régularisation C oubien facteur d'échelle. Ensemble, ils sont connus sous le nom de paramètre d'hyperplan [25]. Pendant la phase de formation, SVM construit un modèle, cartographie la limite de décision pour chaque classe et spécifie l'hyperplan qui sépare les différentes classes. L'augmentation de la distance entre les classes en augmentant la marge de l'hyperplan permet d'augmenter la précision de la classification. Les SVM peuvent également être utilisés pour effectuer efficacement une classification non linéaire [26].

Les SVM ont été appliqués avec succès dans de nombreux domaines divers, notamment la prédiction.

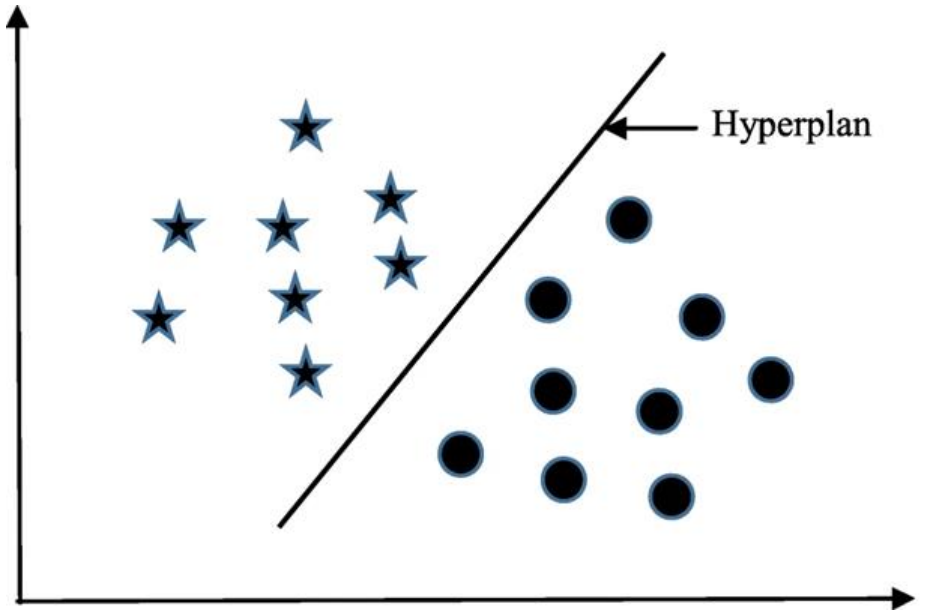


Figure II. 5 Une illustration simplifiée du fonctionnement de la machine à vecteurs de support. Le SVM a identifié un hyperplan (en fait une ligne) qui maximise la séparation entre les classes « étoile » et « cercle ».

Il se résume suivant cet algorithme [26] :

Algorithme générique « SVM »

1. Calculer « d(x, xi) » i =1, 2, ….., n ; où d désigne la distance euclidienne entre les points.
2. Disposez les n distances euclidiennes calculées dans un ordre non décroissant.
3. Soit k un entier +ve, prend les k premières distances de cette liste triée.
4. Trouvez les k-points correspondant à ces k-distances.
5. Soit ki le nombre de points appartenant à la ième classe parmi k points soit $k \geq 0$
6. Si ki >kj $\forall$ i $\neq$ j alors mettez x dans la classe i.

2.2.2.1. Avantages de SVM

1. Leur capacité à manipuler de grandes quantités de données et l'obtention des résultats pertinents en pratique.

2. Le faible nombre d'hyper paramètres utilisés par ces méthodes et que les SVM sont bien fondées théoriquement [48].

2.2.2.2. Inconvénients des SVM

1. Leur utilisation des fonctions mathématiques complexes pour la classification des corpus et pour qu'on trouve les meilleurs paramètres, ce type d'algorithmes demande un temps énorme pendant les phases de test [48].

2.2.3. K-Voisin le plus proche (KNN)

L'algorithme KNN est le modèle le plus souvent utilisé pour la classification, bien qu'il puisse également être utilisé pour l'estimation et Prédiction, aussi est une méthode d'apprentissage basée sur les instances utilisée pour classer les objets en fonction de leurs exemples d'apprentissage les plus proches dans l'espace des caractéristiques. Un objet est affecté à la classe la plus courante parmi ses k plus proches voisins, où k est un entier positif [34,48].

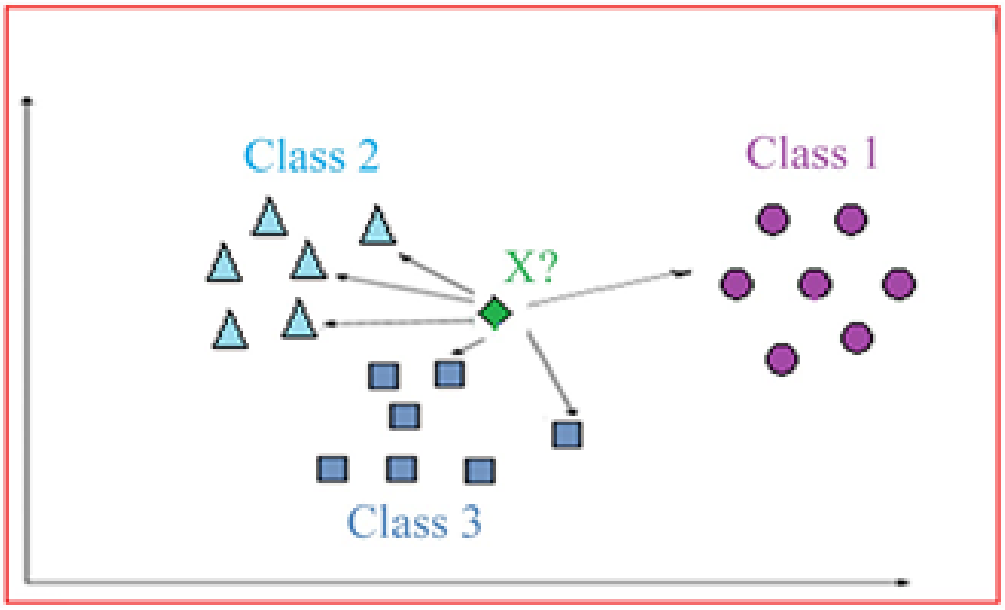


Figure II. 6 Un aperçu illustré simple de l'algorithme K-Nearest Neighbors (KNN).

L'implémentation de l'algorithme KNN se fait en utilisant des métriques de distance euclidiennes pour localiser le voisin le plus proche [35]. Les métriques de distance euclidienne d(x, y) entre deux points x et y est calculé à l'aide de l'équation (4).

$$d(x,y) = \sum_{i=1}^{N} \sqrt{x_i^2 - y_i^2} \qquad (4)$$

où, N est le nombre de caractéristiques telles que, $x=\{x_1,x_2,x_3,\ldots,x_N\}$et $y=\{y_1,y_2,y_3,\ldots,y_N\}$ [19].

Le classificateur KNN est l'une des nombreuses approches qui tentent d'estimer la distribution conditionnelle de Y étant donné X, puis de classer une observation donnée dans la classe avec la probabilité estimée la plus élevée. Étant donné un entier positif K et une observation de test, X0, le classificateur KNN identifie d'abord les points K dans les données d'apprentissage qui sont les plus proches de X0, représenté par N0. Il estime ensuite la probabilité conditionnelle pour la classe j comme la fraction de points dans N0 dont les valeurs de réponse sont égales à j comme indiqué dans l'équation (5) :

$$\Pr(Y = j | X = x_0) = \frac{1}{k} \sum_{i \in N_0} I(y_i = j) \qquad (5)$$

I (yi=j) est une variable indicatrice qui vaut 1 si yi=j et zéro si $y_i \neq j$.

KNN est efficace pour un grand nombre d'exemples de formation et est robuste aux données de formation bruyantes. Mais pour cet algorithme, on détermine en premier lieu la valeur du paramètre k (nombre de plus proches voisins) et le type de distance à utiliser. Le temps de calcul peut être long puisqu'il faut calculer la distance de chaque instance de requête à tous les échantillons d'apprentissage et à mesure que le nombre d'exemples et/ou de prédicteurs/variables indépendantes augmente, il devient beaucoup plus lent [33]. Néanmoins, il n'est pas nécessaire de construire un modèle et de faire des hypothèses supplémentaires ou d'ajuster plusieurs paramètres. KNN est un MLA supervisé polyvalent, simple et facile à mettre en œuvre qui est utilisé pour résoudre des problèmes de de régression, classification et de recherche. L'algorithme suppose que des éléments similaires existent à proximité. Autrement dit, des objets similaires sont proches les uns des autres et que «les oiseaux d'une plume s'assemblent». L'algorithme KNN repose sur le fait que cette hypothèse est suffisamment vraie pour être utile [28].

Le principal inconvénient de KNN est que dans les environnements où les prédictions doivent être faites rapidement, le KNN devenir nettement plus lent à mesure que le volume de données augmente en fait un choix peu pratique. De plus, il existe d'autres algorithmes plus rapides qui peuvent produire des résultats de classification et de régression plus précis. Cependant, à condition qu'il y ait suffisamment de ressources informatiques pour gérer rapidement les données pour faire des prédictions, le KNN peut toujours être utile pour résoudre des problèmes dont les solutions dépendent de l'identification d'objets similaires [36].

Quelques règles sur le choix de k :

Pour sélectionner le K approprié à un ensemble de données, l'algorithme KNN est exécuté plusieurs fois avec différentes valeurs de K et le K qui réduit le nombre d'erreurs rencontrées est choisi tout en maintenant la capacité de l'algorithme à faire des prédictions précises lorsqu'il est appliqué aux données pour lesquelles il n'a pas de contact préalable [37]. Il existe d'autres façons de calculer la distance et une façon peut être préférable selon le problème à résoudre. Cependant, la distance en ligne droite, également appelée distance euclidienne, est un choix populaire et familier [38].

Lorsque la valeur de K diminue jusqu'à 1, les prédictions deviennent moins stables. Inversement, à mesure que la valeur de K augmente, les prédictions deviennent plus stables en

raison du vote à la majorité/de la moyenne, et donc plus susceptibles de faire des prédictions plus précises (jusqu'à un certain point). Finalement, un nombre croissant d'erreurs est constaté. C'est à ce stade que l'on reconnaît que la valeur appropriée de K a été dépassée. La valeur de K est généralement un nombre impair pour avoir un bris d'égalité dans les cas où un vote majoritaire parmi les étiquettes est requis, par exemple, choisir le mode dans un problème de classification [45]. L'algorithme KNN peut être utilisé pour les problèmes de classification, de régression et de recherche. Il est utile pour résoudre des problèmes dont les solutions dépendent de l'identification d'objets similaires.

Il se résume suivant cet algorithme [45] :

Algorithme générique « KNN»

1. Calculer « d(x, xi) » i =1, 2, ….., n ; où d désigne la distance euclidienne entre les points.
2. Disposez les n distances euclidiennes calculées dans un ordre non décroissant.
3. Soit k un entier +ve, prend les k premières distances de cette liste triée.
4. Trouvez les k-points correspondant à ces k-distances.
5. Soit ki le nombre de points appartenant à la ième classe parmi k points soit $k \geq 0$
6. Si ki >kj $\forall$ i ≠ j alors mettez x dans la classe i.

2.2.3.1. Autre avantages de la méthode des k plus proches voisins(KNN)

1. La facilité de mise en œuvre de l'algorithme et
2. Son efficacité pour des classes réparties de manière irrégulière et pour des données incomplètes [48].
3. La méthode des k plus proches voisins n'utilise pas de modèle pour classifier les documents.

2.2.3.2. Autre inconvénients de la méthode des k plus proches voisins(KNN)

1. Pour chaque instance de l'ensemble de donnés on a besoin de calculer la distance ce qui implique un cout de calcul est élevé

2. Sensible aux fonctionnalités non pertinentes [48].

2.2.4. Les forêts aléatoires ou les forêts de décision aléatoires

Sont une méthode d'ensemble pour la classification, la régression et d'autres tâches [48]. Récemment, il y a eu beaucoup d'intérêt pour l'apprentissage d'ensemble, c'est-à-dire les méthodes qui génèrent de nombreux classificateurs et agrègent leurs résultats. Deux méthodes bien connues sont le boosting [39] et le bagging [40] des arbres de classification. Dans le boosting, au cas où les points mal prédits par les prédicteurs précédents, les arbres successifs les donnent un poids supplémentaire. En fin de compte, un vote pondéré est pris pour la prédiction. Dans le bagging, les arbres successifs ne dépendent pas des arbres précédents, chacun est construit indépendamment à l'aide d'un échantillon bootstrap de l'ensemble de données. Au final, un vote à la majorité simple est pris pour la prédiction [29].

Un classificateur RF se compose d'un certain nombre d'arbres, où chaque arbre est développé en utilisant une forme de randomisation (Figure 5). Les nœuds feuilles de chaque arbre sont étiquetés par des estimations de la distribution a posteriori sur les classes d'images. Chaque nœud interne contient un test qui découpe au mieux l'espace des données à classer [32]. Une image est classée en l'envoyant dans chaque arbre et en agrégeant les distributions de feuilles atteintes. Le caractère aléatoire peut être injecté à deux moments de l'apprentissage : lors du sous-échantillonnage des données d'apprentissage afin que chaque arbre soit développé à l'aide d'un sous-ensemble différent, et lors de la sélection des tests de nœuds [30].

Le nombre d'arbres nécessaires à une bonne performance croît avec le nombre de prédicteurs. La meilleure façon de déterminer combien d'arbres sont nécessaires est de comparer les prédictions faites par une forêt aux prédictions faites par un sous-ensemble d'une forêt. Lorsque les sous-ensembles fonctionnent aussi bien que la forêt complète, cela indique qu'il y a suffisamment d'arbres. Pour sélectionner, m try, Breiman [29] suggère d'essayer la valeur par défaut, la moitié de la valeur par défaut et deux fois la valeur par défaut, puis de sélectionner la meilleure. Si l'on a un très grand nombre de variables mais s'attend à ce que très peu d'entre elles soient "importantes", en utilisant un m plus grand, essayez peut donner de meilleures performances. De nombreux arbres sont nécessaires pour obtenir des estimations stables d'importance et de proximité variables. Étant donné que l'algorithme tombe dans la catégorie "parallélisme embarrassant", on peut exécuter plusieurs forêts aléatoires sur différentes machines, puis agréger les composants des votes pour obtenir le résultat final [41].

Le classificateur RF ajoute une couche supplémentaire de caractère aléatoire à l'ensachage [29]. En plus de construire chaque arbre à l'aide d'un échantillon bootstrap différent des données, les RF modifient la façon dont les arbres de classification ou de régression sont construits. Dans les arbres standards, chaque nœud est divisé en utilisant la meilleure répartition parmi toutes les variables tandis que dans un RF, chaque nœud est divisé en utilisant le meilleur parmi un sous-ensemble de prédicteurs de manière aléatoire [42].

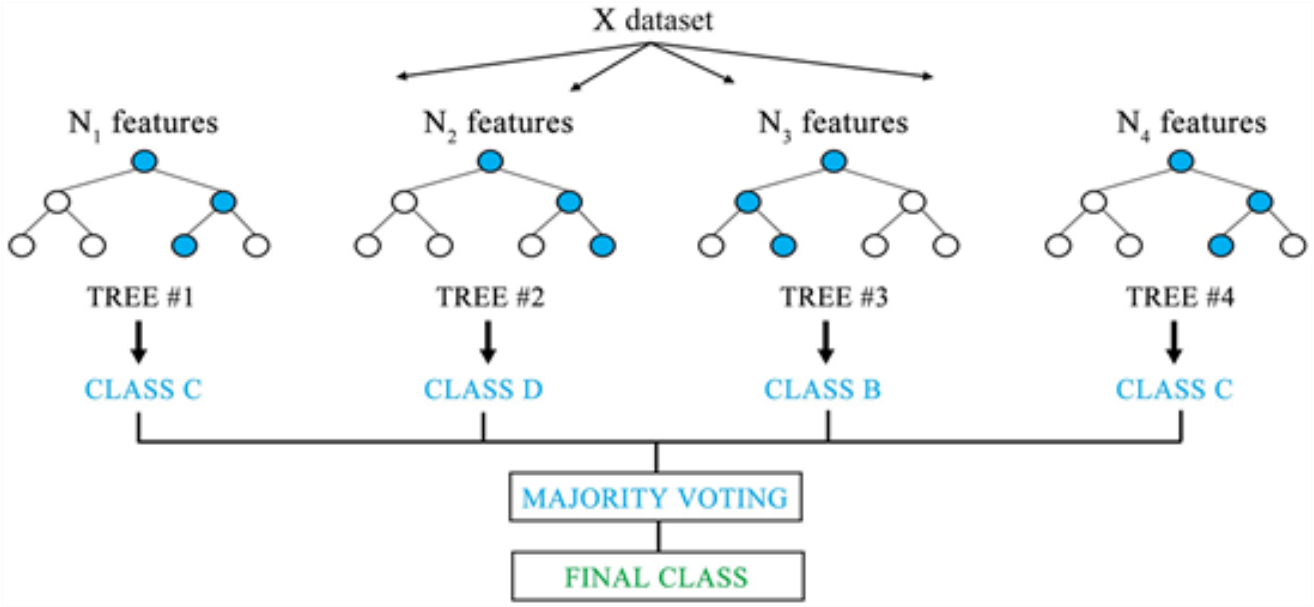


Figure II. 7 Un aperçu illustré de l'algorithme de forêt aléatoire (RF)[49].

Cette stratégie quelque peu contre-intuitive s'avère très performante par rapport à de nombreux autres classificateurs, y compris l'analyse discriminante, les SVM et les NN, et est robuste contre le sur ajustement [27] [29]. De plus, RF est très convivial dans le sens où il n'a que deux paramètres (le nombre de variables dans le sous-ensemble aléatoire à chaque nœud et le nombre d'arbres dans la forêt), et n'est généralement pas très sensible à leurs valeurs [21, 43].

RF est essentiellement un ensemble de DT combinés où chaque arbre vote sur la classe attribuée à un échantillon donné, la réponse la plus fréquente remportant le vote [44]. Cet algorithme peut très bien gérer les caractéristiques catégorielles, peut également gérer des espaces de grande dimension ainsi qu'un grand nombre d'exemples d'apprentissage [29]. Les RF sont assez polyvalents, d'où leur popularité et leur application dans divers domaines. Un arbre de décision est un ensemble de conditions organisées dans une structure hiérarchique. Il s'agit d'un modèle prédictif dans lequel une instance est classée en suivant le chemin des

conditions satisfaites depuis la racine de l'arbre jusqu'à atteindre une feuille, qui correspondra à une étiquette de classe. Une DT peut facilement être convertie en un ensemble de règles de classification [31].

Il se résume suivant cet algorithme [31] :

Algorithme générique « Forêt aléatoire »

1. ***Entrées :*** L'ensemble d'apprentissage L, Nombre d'arbres N.
2. ***Sortie :*** Ensemble d'arbres E
3. ***Processus :*** for i = 1 → N do
4. Ti ← BootstrapSample(T)
5. Ci ← ConstructTree(Ti) où à chaque nœud :
6. – Sélection aléatoire de K = √ M Variables à partir de l'ensemble d'attributs M
7. – Sélection de la variable la plus informative K en utilisant l'index de Gini
8. – Création d'un nœud fils en utilisant cette variable
9. E ← E{C i}
10. *end for*
11. Retourner E

2.2.5. La méthode de l'arbre de décision

L'arbre de décision représente une méthode très efficace d'apprentissage supervisé et un outil de modélisation prédictive qui peut être appliqué dans de nombreux domaines, consiste à organiser les données étudiées sous forme d'arbre avec une racine, des branches et des feuilles. Généralement le processus récursif est utilisé pour obtenir un résultat optimum de la valeur de la classe faisant l'objet de la prédiction [47,48].

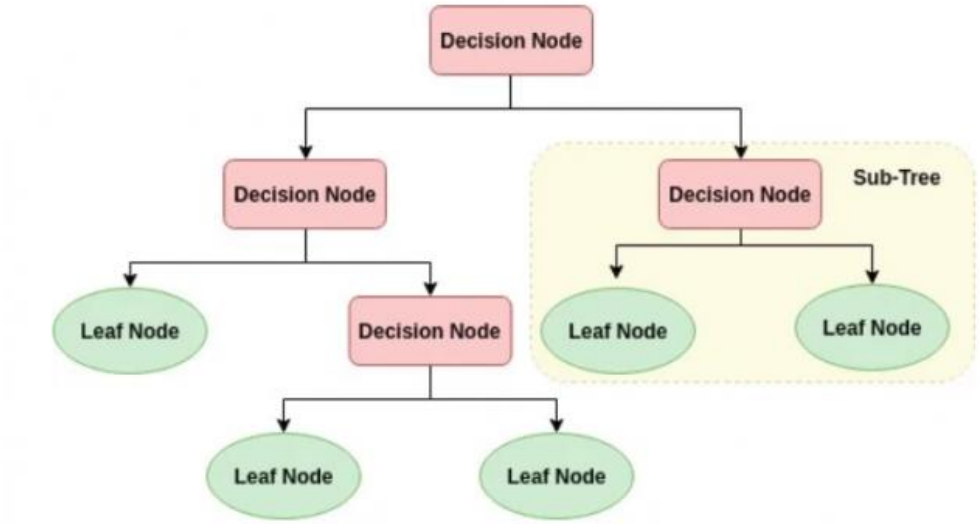


Figure II. 8 Schéma simplifié d'un arbre de décision.

Il se résume suivant cet algorithme [54] :

Algorithme générique de construction d'un arbre de décision

1. Initialiser l'arbre courant à l'arbre vide ; la racine est le nœud courant
2. **Répéter**
3. Décider si le nœud courant est terminal (une feuille).
4. **Si** le nœud est terminal **Alors**
5. Affecter une classe au nœud courant
6. **Sinon**
7. Sélectionner un test et créer autant de nouveaux fils qu'il y a de possibilités de réponses
8. **Finsi**
9. Passer au nœud suivant
10. **Jusqu'à** l'obtention de l'arbre de d´décision

2.2.5.1. Avantage des arbres de décision

1. Leur capacité à travailler sur des données symboliques et la facilité d'apprentissage et d'utilisation.

2. Leur grande capacité et efficacité à faire de la classification [48].

2.2.5.2. Inconvénients des arbres de décision

1. Ce sont des algorithmes très sensibles aux points aberrants et au bruit et au changement des données [48].

2.3. Etude comparative entre les algorithmes prédictifs de SML

Tous les techniques d'apprentissage automatique supervisé citer dans la revue de la littérature sont applicables dans de nombreux domaines notamment la prédiction et s'adaptent à des situations non linéaires [50 ,51].

Cette étude a conclu que les SVM a tendance à être beaucoup plus performants lorsqu'ils traitent de plusieurs dimensions et de fonctionnalités continues. Ainsi que, pour le modèle SVM, une grande taille d'échantillon est nécessaire pour atteindre sa précision de prédiction maximale et que les SVM fonctionnent bien lorsque la multi-colinéarité est présente et qu'une relation non linéaire existe entre les caractéristiques d'entrée et de sortie. Aussi, les algorithmes KNN, random Forest et l'arbre de décision apprend des interactions pertinentes parmi les prédicteurs par rapport au NB et SVM.

La plupart des algorithmes d'arbre de décision ne peuvent pas bien fonctionner avec des problèmes qui nécessitent un partitionnement diagonal et que les forets aléatoires ont de meilleures performances prédictives que les arbres de décision. Ainsi qu'il est généralement admis que k-NN est très sensible aux caractéristiques non pertinentes : cette caractéristique peut s'expliquer par le fonctionnement de l'algorithme.

Naïve Bayes (NB) fait des calculs rapides sur des jeux de données volumineux et nécessite peu d'espace de stockage pendant les phases d'apprentissage et de classification : le strict minimum est la mémoire nécessaire pour stocker les probabilités a priori et conditionnelles. L'algorithme kNN de base utilise beaucoup d'espace de stockage pour la phase d'apprentissage, et son espace d'exécution est au moins aussi grand que son espace d'apprentissage. Au contraire, pour tous les apprenants non paresseux, l'espace d'exécution est généralement beaucoup plus petit que l'espace d'entraînement, car le classificateur résultant est généralement un résumé très condensé des données. De plus, Naïve Bayes et le kNN peuvent être facilement utilisés comme apprenants incrémentaux alors que les algorithmes de règles ne le peuvent pas. Naïve Bayes est naturellement robuste aux valeurs manquantes puisque celles-ci sont simplement ignorées dans le calcul des probabilités et n'ont donc aucun impact sur la décision finale. Au contraire, kNN et les réseaux de neurones nécessitent des enregistrements complets pour faire leur travail.

Enfin, les arbres de décision et les NB ont généralement des profils opérationnels différents, lorsque l'un est très précis, l'autre ne l'est pas et vice versa. Au contraire, les arbres de décision et les classificateurs de règles ont un profil opérationnel similaire. SVM ont également un profil opérationnel similaire. Aucun algorithme d'apprentissage ne peut surpasser uniformément les autres algorithmes sur tous les ensembles de données.

Différents ensembles de données avec différents types des variables et le nombre d'instances déterminant le type d'algorithme qui fonctionnera bien. Il n'y a pas d'algorithme d'apprentissage unique qui surpassera les autres algorithmes basés sur tous les ensembles de données selon le théorème d'absence de repas gratuits [50,52].

2.4. Conclusion

Les techniques d'apprentissage supervisé ont obtenu un grand succès dans le domaine de la prédiction.

Dans ce chapitre nous avons essayé d'exposer les différentes méthodes prédictives d'apprentissage automatique supervisé à savoir : l'arbre de décision, Random Forest, naïve bayes, K nearest Neighbors et support vector machine, et faire une comparaison entre eux.

Dans le prochain chapitre nous abordons la méthode de Monte Carlo qui appartient aux algorithmes probabilistes « Naïve Bayes » d'apprentissage automatique supervisé et l'application de cette méthode afin de résoudre notre problème.

Chapitre 3 :
Système de prédiction proposé

Chapitre 3 : Le système de prédiction proposé.

3.1. Introduction

Dans toute compétition sportive, il y a un grand intérêt à savoir quelle équipe sera le champion à la fin du championnat c'est le cas de football, dont les pronostics intéressent beaucoup les fans et la presse sportive. Ces dernières années, c'est l'objet de plusieurs études. Dont beaucoup de développeurs s'intéressent de prédiction de résultat final des matchs de la coupe du monde, dans notre algorithme de prédiction on a basé sur la loi de poisson et la méthode Monte Carlo qui appartient aux algorithmes probabilistes « Naïve Bayes » d'apprentissage automatique supervisé.

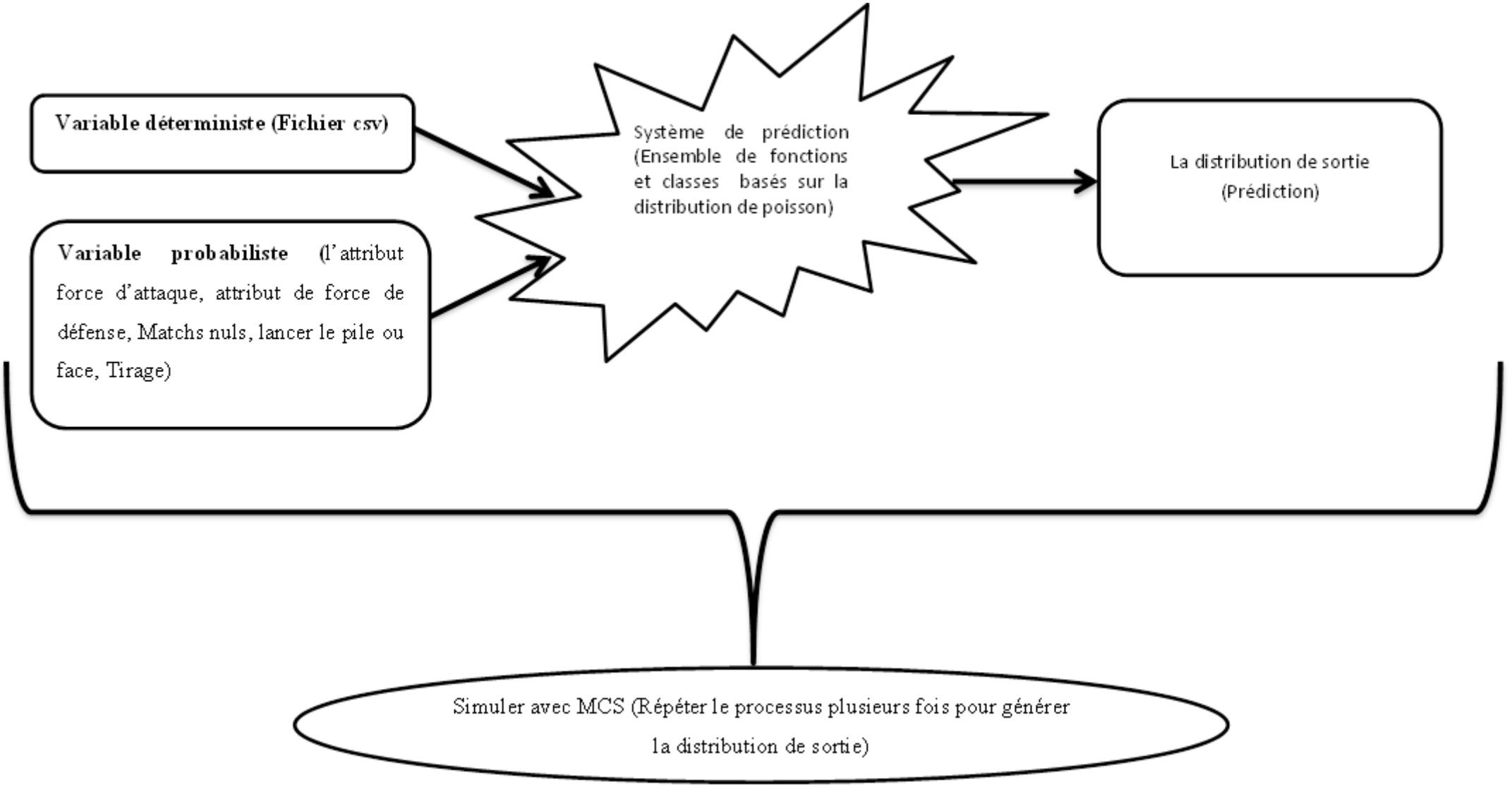
De nos jours, la méthode de Monte Carlo est utilisée dans un vaste ensemble de disciplines, parmi ces dernières on se base sur la branche des mathématiques qui s'intéresse aux expériences sur des nombres aléatoires.

Lorsque la résolution mathématique d'un problème donné n'est pas possible, on fait appel à des méthodes d'approximation. Cela permet de modéliser des situations très complexes dont la solution analytique s'avère difficile voire impossible. Parmi ces méthodes d'approximation figure la simulation, qui représente un outil très utile, elle permet de valider ou d'invalider des hypothèses, d'obtenir des informations quantitatives, où tout simplement d'explorer le comportement d'un modèle lorsque celui-ci est mal connu où mal compris [55].

Dans ce chapitre, nous allons présenter quelques notions concernant la simulation, plus particulièrement, la simulation de Monte Carlo, ainsi que les techniques mathématiques sous-jacentes requises pour effectuer cette simulation cependant les bases de cette méthode et de son utilisation pour la prédiction.

- En premier lieu, on commence par définir les concepts généraux de la simulation.
- Puis, on passe à la définition du concept global de la coupe du monde et la définition globale de notre algorithme de prédiction.

3.2. Conception de système de prédiction de vainqueur de la coupe du monde



3.3. La simulation

La simulation est une technique numérique pour conduire des expériences sur un ordinateur qui peut inclure des caractéristiques stochastiques et implique l'utilisation de modèles mathématiques qui décrivent le comportement d'un système [59].

3.3.1. La simulation Monte Carlo

3.3.1.1. Définition générale

Vers les années 1949 la méthode de Monte Carlo a été développée mathématiciens américains John Von Neumann et Stanislav Ulam, ce n'est toutefois qu'avec l'avènement des ordinateurs que l'on a pu réellement l'utiliser. Cette méthode est une famille de méthodes algorithmiques visant à calculer une valeur numérique approchée en utilisant des procédés aléatoires. Quant au nom de Monte Carlo, on le doit bien sûr à la capitale de la principauté de Monaco.

L'expression "Simulation de Monte-Carlo" consiste en l'utilisation du hasard pour aborder l'étude d'un problème déterministe. Elle est quasiment incontournable dans l'approche de certains problèmes tels que la complexité de la résolution par voie purement mathématique et lorsque le problème est trop volumineux (en particulier, contient un trop grand nombre de variables) pour que les techniques d'approximation numérique puissent conduire à un résultat précis dans un temps acceptable. Ce genre de méthode compte sans doute parmi les méthodes les plus utilisées dans tous les domaines ayant recours aux mathématiques appliquées : physique, chimie, biologie, économie, sociologie etc... En prends à titre d'exemple La prédiction [56].

L'algorithme de Monte Carlo est utilisé pour une multitude d'applications. Ils sont utilisés de manière routinière et intensive dans "la finance, l'ingénierie, la chaîne d'approvisionnement et la science"[57]. L'appellation Monte Carlo fait référence aux jeux du casino de monte Carlo auxquels s'apparente cette méthode puisqu'elle fait appel à des nombres aléatoires.

La simulation de Monte Carlo de la "réalité" à travers un certain nombre de mesures aléatoire simple est souvent le seul moyen réaliste de la réalisation des calculs dans les grands systèmes complexes.

3.3.1.2. Sur le Choix de la méthode

L'usage de l'approche Monte-Carlo dans le cadre de l'élaboration de la prédiction des matchs de la coupe du monde va vous permettre de réduire considérablement l'incertitude entourant vos hypothèses et de mesurer plus précisément le risque rattaché à certaines des variables de votre modèle d'affaires. Ainsi, vous gagnerez confiance en vos prévisions et vous maîtriserez mieux les éléments d'incertitude de système.

Ainsi que La simulation Monte Carlo présente plusieurs avantages tels que :

- **Résultats probabilistes :** Les résultats indiquent non seulement ce qui pourrait arriver, mais dans quelle mesure ;

- **Résultats graphiques :** Les données produites par la simulation Monte Carlo facilitent la représentation graphique des différentes issues et de leur chance de se produire. La présentation des conclusions de l'analyse en est d'autant plus simple ;

- **Analyse de sensibilité :** Dans la simulation Monte Carlo, les entrées qui produisent le plus d'effet sur les résultats se distinguent clairement ;

- **Analyse de scénario :** Avec la simulation Monte Carlo, l'analyste voit clairement les combinaisons de valeurs en entrée associées aux issues et dispose ainsi d'une information extrêmement utile à la poursuite de l'analyse ;

- **Corrélation des entrées :** Dans la simulation Monte Carlo, il est possible de modéliser des rapports interdépendants entre les variables en entrée. Il est en effet important de représenter, pour la précision du modèle, la manière dont la hausse de certains facteurs s'accompagne dans la réalité de celle d'autres facteurs ou, au contraire, de leur baisse [58].

3.3.2. La démarche suivie

Partie 1 :

La Coupe du Monde de la FIFA, c'est le plus grand tournoi de football au monde dans lequel 32 nations s'affrontent pour la coupe du monde. Les équipes sont choisies sur la base de matches amicaux internationaux qui ont lieu avant le début de la coupe du monde. Les 32 meilleures équipes qualifiées sont classées en 8 groupes et s'affrontent dans leurs phases de groupes respectives.

Les 2 meilleures équipes de chaque groupe entrent en huitièmes de finale. Cela marque la phase à élimination directe du tournoi. Les quarts de finale, les demi-finales et la finale sont les tours à suivre. Chaque tour élimine des équipes et les équipes gagnantes passent au tour suivant. En fin de compte, deux meilleures équipes se battent pour le trône de la Coupe du monde.

Partie 2 :

Il existe de nombreux modèles sophistiqués que les gens peuvent construire pour résoudre un problème de prédiction. Dans notre cas On a simulé avec la méthode de Monte Carlo qui est un type de simulation qui exige que le même événement se reproduise plusieurs fois indépendamment pour conclure la probabilité de tous les résultats aussi précisément que possible.

Et à travers nos recherches sur la méthode de Monte Carlo (MC) on a conclu que c'est un outil mathématique qui peut être appliquée dans différents domaines de la science de l'ingénieur. C'est une approche probabiliste permettant la modélisation des paramètres incertains dans des systèmes compliqués. L'algorithme fondamental de la méthode de Monte Carlo se résumé en trois phases.

- Définir un domaine des entrées possibles.
- Produire des entrées aléatoirement du domaine, et exécutez un calcul déterministe sur elles.
- Agrégez les résultats des différents calculs dans le résultat final.

Le but de cet algorithme est de prédire le vainqueur de la coupe du monde de la FIFA, ainsi que la probabilité de chaque équipe de gagner la coupe du monde. Nous avons décidé de prédire les vainqueurs des matchs de la coupe du monde.

Phase 1 : Définir un domaine des entrées possibles

Lancer la pile ou face

Le lancer est un facteur crucial à prendre en compte pour un match joué. Nous pensons que TOSS pourrait jouer un rôle vital dans le résultat d'un match. Notre modèle randomise le Toss entre deux équipes jouant un match. L'équipe qui remporte le tirage au sort bénéficie d'un léger avantage en termes de probabilité selon notre modèle. Nous pensons que l'équipe qui remporte le tirage au sort bénéficie d'un avantage stratégique car elle peut choisir entre

deux options : sélection du côté et coup d'envoi Nous avons analysé la tendance au fil des ans et observé que les équipes qui remportent le tirage au sort ont un meilleur taux de réussite.

Attribut force d'attaque

Cette variable est utilisée pour démontrer la force d'attaque de l'équipe, elle est généré aléatoirement en fonction du classement de l'équipe. Cela peut être un facteur majeur pour décider du nombre de buts que l'équipe marque. On trouve d'abord la moyenne des buts marqués par une équipe grâce à cette variable. Enfin, nous calculons l'attaque moyenne de l'équipe.

Attribut force de défense

Cette variable est utilisée pour démontrer la puissance de défense de l'équipe, elle est généré aléatoirement en fonction du classement de l'équipe. Cela peut être un facteur majeur pour décider du nombre de buts que l'équipe concède. On trouve d'abord la moyenne de buts encaissés par une équipe grâce à cette variable. Enfin, nous calculons la défense moyenne de l'équipe.

Tirages

Nous randomisons les tirages pour chaque tour. Ainsi, chaque tour contient des groupes avec différentes équipes réparties au hasard. Cette variable est générée aléatoirement pour éviter de définir des paramètres pour décider quelle équipe va dans quel groupe. Toutes les coupes du monde ont des créneaux prédéfinis pour les équipes. Mais notre modèle inclut cette fonctionnalité innovante qui rend cela aléatoire et supprime donc toute sorte de biais. C'est une sorte de suspense que les équipes s'affronteront dans les tours pour continuer.

Les groupes seront les suivants, tels que Toutes les équipes sont indépendantes les unes des autres, n'importe quelle équipe peut être affectée à n'importe quel groupe et Les performances de chaque équipe sont indépendantes les unes des autres:

Phase de groups

- Groupe A: équipe 1, équipe 2, équipe 3, équipe 4
- Groupe B: équipe 5, équipe 6, équipe 7, équipe 8

- Groupe C: équipe 9, équipe 10, équipe 11, équipe 12
- Groupe D: équipe 13, équipe 14, équipe 15, équipe 16
- Groupe E: équipe 17, équipe 18, équipe 19, équipe 20
- Groupe F: équipe 21, équipe 22, équipe 23, équipe 24
- Groupe G: équipe 25, équipe 26, équipe 27, équipe 28
- Groupe H: équipe 29, équipe 30, équipe 31, équipe 32

Phase à élimination directe

- **HUITIEME DE FINALE:**
  - Quart-Finaliste1: Equipe-qualifié 1 vs Equipe-qualifié 2
  - Quart-Finaliste2: Equipe-qualifié 3 vs Equipe-qualifié 4
  - Quart-Finaliste3: Equipe-qualifié 5 vs Equipe-qualifié 6
  - Quart-Finaliste4: Equipe-qualifié 7 vs Equipe-qualifié 8
  - Quart-Finaliste5: Equipe-qualifié 9 vs Equipe-qualifié 10
  - Quart-Finaliste6: Equipe-qualifié 11 vs Equipe-qualifié 12
  - Quart-Finaliste7: Equipe-qualifié 13 vs Equipe-qualifié 14
  - Quart-Finaliste8: Equipe-qualifié 15 vs Equipe-qualifié 16
- **QUART FINALE:**
  - Demi-Finaliste1: Equipe-qualifié 1 vs Equipe-qualifié 2
  - Demi-Finaliste2: Equipe-qualifié 3 vs Equipe-qualifié 4
  - Demi-Finaliste3: Equipe-qualifié 5 vs Equipe-qualifié 6
  - Demi-Finaliste4: Equipe-qualifié 7 vs Equipe-qualifié 8
- **DEMI FINALE:**
  - Finaliste1: Equipe-qualifié 1 vs Equipe-qualifié 2
  - Finaliste2: Equipe-qualifié 3 vs Equipe-qualifié 4
- **FINALE**:
  - Finaliste1 vs Finaliste2

Phase 2 : Produire des entrées aléatoirement du domaine, et exécutez un calcul déterministe sur elles

Dans cette étape On utilise Les entrées (Lancer la pile ou face, Tirages, force d'attaque, force de défense) pour déterminer une sortie à base de la distribution de poisson.

Avant de commencer il faut calculer ces valeurs à base de fichier csv , tels que les données de ce dernier se sont des données du monde réel et du site Web officiel de la coupe du monde de la FIFA, qui comprend le classement des équipes, les performances passées de l'équipe dans les coupes du monde (nombre de finales gagnées et le nombre de matchs jouées) :

- MoyButsmarqué = Sum(Attaque)/len(équipes)
- MoyButsencaissé = Sum(Defense)/len (équipes)
- tauxréussite = nbrfinalgagnées / nbrmatchsjoués
- AttaqueMoy =Attaque/ MoyButsmarqué + tauxréussite
- DefenseMoy = Defense / MoyButsencaissé + tauxréussite

La fonction Toss_factor : Cette fonction calcule le tirage au sort au début de chaque match. Notre modèle suppose que l'équipe qui remporte le tirage au sort a un léger avantage sur l'adversaire.

La fonction calcule aléatoirement le gagnant du tirage au sort. Sur la base de ce résultat, nous attribuons un nombre généré aléatoirement entre 0,5 et 1 à l'équipe perdante et [1- (ce nombre)] à l'équipe gagnante. Nous le faisons parce que nous allons multiplier ce nombre (Facteurlancer) par le dénominateur d'une équation utilisée dans la fonction match_between.

Class WorldCupTeam: Cette classe attribue des valeurs d'attaque et de défense à chaque équipe qui peuvent être utilisées par la fonction match between pour calculer le résultat.

Buts marqués par l'équipe 1 ou 2

Cette variable est la représentation réelle des buts marqués dans le match par l'équipe 1 ou 2. Elle est générée aléatoirement à l'aide de la distribution de Poisson. Nous utilisons la distribution de Poisson sur (moyenne de buts marqué par) pour générer aléatoirement un score pour ce match par équipe 1 ou 2. Nous utilisons cette variable pour faire la distinction entre le gagnant et le perdant du match. Sur la base de cette variable, nous attribuons des points, des buts marqués et une différence de buts pour les phases de groupe.

MoyButsEquipe 1 = AttaqueMoy (équipe1)/ DefenseMoy (équipe2) * Facteurlancer 1

MoyButsEquipe2 = AttaqueMoy (équipe2)/ DefenseMoy (équipe1) * Facteurlancer 2

La fonction Match_between : Cette fonction est la plus importante du programme. Il calcule le résultat d'un match entre deux équipes et la probabilité que les deux équipes gagnent où la probabilité d'égalité entre les deux équipes.

C'est le modèle qui génère aléatoirement le score du match en utilisant la distribution de Poissons et utilise Facteurlancer de l'équipe 1 et le Facteurlancer de l'équipe 2 de la fonction Toss_factor et Buts marqués par l'équipe 1 ou 2.

score_équipe1 = np.random.poisson (MoyButsEquipe 1)

score_équipe2 = np.random.poisson(MoyButsEquipe2)

La fonction Qualifiedteams: Cette fonction trouve les deux meilleures équipes qualifiées de chaque groupe. Les deux meilleures équipes de chaque groupe seront choisies en fonction de leurs « points », « différence de buts », « score_équipe »

La fonction Main : C'est la fonction principale du programme. Les utilisateurs ont le choix de simuler soit des tirages fixes, soit des tirages aléatoires pour l'ensemble du tournoi.

Nous avons incorporé des tirages fixes et des tirages aléatoires

Des tirages fixes : nous attribuions les valeurs d'attaque et de défense à chaque équipe en fonction du classement de leur équipe. Nous avons créé des bacs de taille 8 et séparé 32 équipes en 4 bacs de ce type en fonction de leur classement. Nous avons ensuite généré aléatoirement des valeurs d'attaque et de défense pour les équipes par rapport à ces bacs.

Des tirages aléatoires : pour les équipes de premier ordre, c'est-à-dire les 8 meilleures équipes auront des valeurs d'attaque et de défense dans une gamme et ainsi de suite. Cela a réduit toute sorte de biais et amélioré l'efficacité. Notre modèle randomise les tirages, les scores et les lancers à chaque fois.

Phase 3 : Agrégez les résultats des différents calculs dans le résultat final

A la fin de chaque simulation on obtient ce qui gagne la finale et on le stocke et après les 10000 simulations, l'équipe qui a la plus grande probabilité de gagner sera le vainqueur prédit et les positions peuvent être décidées en fonction de la probabilité de victoire des autres équipes.

3.4. Résumé de la méthode

Nous allons randomiser les tirages après chaque tour. Nous trions d'abord les équipes en fonction de leur classement FIFA, puis nous leur attribuons au hasard 2 valeurs - Attaque, Défense en fonction de leur classement. Le vainqueur de chaque phase de groupes est déterminé à partir (des points obtenus par l'équipe, des buts marqués et de la différence de buts). Nous avons également envisagé de donner un avantage à l'équipe qui remporte un tirage au sort (il sera complètement randomisé pour éviter les biais). Pour une précision optimale, notre modèle utilise une équation qui effectue des calculs sur ces valeurs d'ATTAQUE ET DE DÉFENSE générées aléatoirement et nous donne des valeurs mises à jour. Nous essayons également d'intégrer les performances passées des équipes lors des Coupes du monde (rapport de victoires).

Dans la phase de groupes, l'équipe gagnante obtient 3 points. En cas d'égalité lors de la phase de groupes, les deux équipes obtiennent 1 point. Cependant, si nous rencontrons un match nul dans la phase à élimination directe, nous le traitons comme un ET (Extra Time). Ainsi, jusqu'à ce que le score soit inégal, notre modèle simule le résultat final du match.

Nous prédisons la moyenne des buts marqués par l'équipe 1 contre l'équipe 2 en utilisant les équations suivantes :

MoyButsEquipe 1 = AttaqueMoy (équipe1)/ DefenseMoy (équipe2) * Facteurlancer 1

MoyButsEquipe2 = AttaqueMoy (équipe2)/ DefenseMoy (équipe1) * Facteurlancer 2

Après plusieurs discussions, nous avons trouvé un moyen plus efficace d'attribuer une valeur d'attaque et de défense à chaque équipe. Auparavant, nous attribuions ces valeurs en fonction du classement de leur équipe (Simulations aléatoire). Nous avons ensuite créé des bacs de taille 8 et séparé 32 équipes en 4 bacs de ce type en fonction de leur classement (Simulations fixes). Nous avons ensuite généré aléatoirement des valeurs d'attaque et de défense pour les équipes par rapport à ces bacs. Alors maintenant, pour les équipes de premier ordre, c'est-à-dire les 8 meilleures équipes auront des valeurs d'attaque et de défense dans une gamme et ainsi de suite. Cela a réduit toute sorte de biais et amélioré l'efficacité. Notre modèle randomise les tirages, les scores et les lancers de pile ou face à chaque fois et nous n'avons donc pas pu incorporer plusieurs Doctests. Cependant, nous avons inclus 1 doctest en définissant la graine pour obtenir la même sortie à chaque fois. Cela peut être simplement

utilisé comme exemple sur quelle entrée donné et comment sera notre sortie. Nous avons également créé une fonction TOSS distincte qui peut être utilisée avant chaque match. Toujours dans le respect des suggestions de nos pairs, nous avons incorporé des tirages fixes et des tirages aléatoires. Notre modèle donne la flexibilité à l'utilisateur de choisir s'il veut des tirages fixes ou des tirages aléatoires. Les tirages fixes contiennent des combinaisons d'équipes que la Coupe du Monde de la FIFA 2022 à prédécédés ainsi que les équipes d'autres coupes du monde dans le cas de la prédiction du vainqueur d'une coupe du monde choisie . Les tirages au sort seront notre approche principale qui randomisera les groupes et les équipes. Nous avons prédit avec succès la probabilité de chaque équipe de gagner la Coupe du monde après avoir simulé le tournoi 10 000 fois par la méthode de monte Carlo qui est une méthode statistique dans laquelle de nombreuses expériences aléatoires similaires sont réalisées.

On va suivre un modèle de simulation stochastiques La simulation stochastique s'applique au processus dites : 'Stochastique' ou processus 'Aléatoire', qui représentent une évolution, généralement dans le temps, d'une (des) variable(s) aléatoire(s). Au contraire, un modèle de simulation stochastique nécessitera la connaissance de lois de probabilité pour représenter le système et une simulation du hasard pour décrire son fonctionnement. La plupart des systèmes sont stochastiques et l'on approche souvent leur comportement en faisant intervenir des lois de probabilité [58].

Nous avons d'abord simulé une correspondance unique basée sur les paramètres ci-dessus. Ensuite, nous avons simulé la phase de groupes. En fin de compte, nous simulons la phase à élimination directe pour découvrir le gagnant. Au cours de cette simulation, nous avons pensé que la précision du modèle pouvait être augmentée si nous utilisions le taux de victoire des équipes lors des coupes du monde précédentes. Nous avons donc intégré cela dans notre modèle.

Nous utilisons la distribution de Poisson pour deux raisons principales.

Premièrement, la distribution de Poisson fonctionne mieux lorsque les variables sont indépendantes les unes des autres. Donc, en supposant ici que la moyenne des buts marqués par équipe est indépendante, nous avons décidé d'utiliser la distribution de Poisson.

Deuxièmement, la distribution de Poisson fonctionne mieux lorsque nous connaissons un seul paramètre (buts moyens marqués par équipe) dans notre cas.

Sachant que la distribution de Poisson est bien adaptée aux conditions d'objectif car elle montre la "distribution des événements rares". La probabilité qu'un certain ratio de buts se produise dans un match de football est très faible en raison du grand nombre de résultats possibles.

3.5. Conclusion

Pour prédire l'issue des matchs de la coupe du monde et obtenir des résultats de prédictions du niveau d'un être humain qui suit le Football régulièrement et connaît bien la plupart des équipes, nous avons utilisé la méthode de monte Carlo qui est une méthode probabiliste.

Dans ce chapitre, nous avons expliqué le contexte global de notre système de prédiction qui est basé essentiellement sur la méthode de Monte Carlo et on a définis quelques fonctions et classes de l'algorithme.

Notre modèle donne la flexibilité à l'utilisateur de choisir s'il veut des tirages fixes ou des tirages aléatoires. Les tirages fixes contiennent des combinaisons d'équipes que la Coupe du Monde de la FIFA 2022 à prédécédés ainsi que les autres coupes du monde. Les tirages au sort seront notre approche principale qui randomisera les groupes et les équipes. Nous avons prédit avec succès la probabilité de chaque équipe de gagner la Coupe du monde après avoir simulé le tournoi 10 000 fois par la méthode de monte Carlo qui est une méthode statistique dans laquelle de nombreuses expériences aléatoires similaires sont réalisées.

Dans le chapitre suivant, nous allons présenter l'aspect pratique de notre application ainsi que les différents outils et logiciels utilisés tout au long de développement de l'application.

Chapitre 4 : Réalisation

Chapitre 4 : Réalisation.

4.1. Introduction

Nous arrivons dans ce chapitre à la description de l'aspect pratique de notre travail. Dans la description de notre plateforme qui suivra, nous mettrons l'accent sur le côté visuel (les interfaces) afin montrer sa facilité d'utilisation qui nous a été un objectif principal. En effet, Nous avons essayé de concevoir une interface intuitive et pratique. Nous décrivons aussi dans ce chapitre l'ensemble des moyens technologiques utilisés dans le développement de notre plateforme. Rappelons que, le projet que nous décrivons à travers ce rapport concerne la réalisation d'un système de prédiction de vainqueur de la coupe du monde

Tout développement de projet informatique nécessite le choix des technologies adéquates à son implémentation. Et c'est en présentons lieu la plateforme de développement de notre application et le format du fichier de stockage de la base de données, ainsi que certains que nous avons utilisé et la manière dont notre système a été élaboré que nous débutons ce dernier chapitre. Par la suite, nous allons présenter les interfaces de notre application desktop afin de mettre en évidence leurs aspects pratiques et intuitifs qui nous ont été l'un de nos principaux objectifs. Rappelons que notre projet consiste à réaliser un système de prédiction du vainqueur de la coupe de monde ainsi que l'issue de chaque match.

4.2. Les ressources logicielles :

4.2.1. L'environnement de développement (IDE)

Il y a plusieurs IDE, certaines sont polyvalent, d'autre sont utilisés pour des systèmes d'exploitation spécifiques. Parmi ces IDE le plus utilisé est Spyder.

4.2.1.1. Spyder

Spyder est un environnement écrit en Python pour Python, conçu essentiellement pour des analyses de données. Il permet d'éditer, d'exécuter et de déboguer du code dans un seul environnement. Il présente un outil de développement complet avec l'exploration de données, l'exécution interactive, l'inspection approfondie et les belles capacités de visualisation d'un package scientifique [34].

4.2.3. Langage de programmation

4.2.3.1. Présentation de langage Python

Est un langage de programmation le plus populaire et le plus utilisé dans le domaine du Machine Learning, du Big Data et de la Data Science. Aussi un langage très accessible, qui facilite plusieurs tâches axées sur la programmation, puisqu'il existe des centaines de milliers du packages pythons disponibles qui fournissent des modules ainsi que des outils nécessaires pour les fonctionnalités courantes [60].

4.2.3.2. Versions de Python

On distingue deux versions de python :

Python 2.x : est une version legacy, qui continuera d'être supportée et de recevoir des mises à jour officielles jusqu'en 2020. Et Après, elle continuera d'ailleurs sans doute de subsister de façon non officielle [62].

Python 3.x : est la version courante, elle apporte de nouvelles fonctionnalités et très utiles**,** telles qu'un meilleur contrôle de concurrence et un interpréteur plus efficace, c'est la version qui remplace de plus en plus la version 2. Et l'adoption de Python 3 a été ralentie par le manque de bibliothèques tierces prises en charge, puisqu'un grand nombre d'entre elles n'étaient compatibles qu'avec Python 2, ce qui compliquait la transition. Ce problème est aujourd'hui pratiquement résolu et il reste peu de raisons valables de continuer à utiliser Python 2 [62].

4.2.3.3. Contexte d'utilisation

La Data Science : il est utilisé dans les data science puisque est un langage simple, lisible, propre, flexible et compatible avec de nombreuses plateformes. Aussi les nombreuses bibliothèques, telles que TensorFlow, Scipy et Numpy permettent d'effectuer plusieurs tâches [62].

Programmation système : Python est utilisé comme langage de script système par défaut**.**

Le domaine scientifique : Est utilisé pour des librairies scientifiques « classiques » portées ou appelées, interface pour code métier.

Le web : Pour développer des sites et leurs interfaces (on prend comme exemple « Django »).

Développement logiciel : Développer l'interface de Scripting pour des codes plus bas niveaux.

Surcouche logicielle : python est un outil de développement d'add-on sur application préexistante, langage de liaison entre différente base logicielle [63].

4.2.3.4. Avantages

- **Bibliothèques standard très riche :** Le python a une bibliothèque complète et contient du code à des fins diverses.
- **Syntaxe très claire :** est un langage facile à lire, apprendre, comprendre et à coder. Il n'a pas non plus besoin d'accolades pour définir les blocs, et l'indentation est obligatoire. Cela facilite davantage la lisibilité du code.
- Facile à apprendre tout en conservant une flexibilité de paradigme (procédural, fonctionnel ou POO).
- Multi-plateforme.
- **Libre et open-source :** Python est disponible gratuitement ainsi que son code source est gratuit, y apporter des modifications et même le distribuer. Il se télécharge avec une vaste collection de bibliothèques pour vous aider dans vos tâches [64].

4.2.3.5. Inconvénients

Limitations de vitesse : l'exécution python est faite ligne par ligne. Mais comme Ce langage est interprété, il en résulte souvent exécution lente. Cependant, à moins qu'une vitesse élevée ne soit requise, les avantages offerts par Python sont suffisants pour nous distraire de ses limitations de vitesse.

Faible dans l'informatique mobile et les navigateurs : Est un excellent langage côté serveur contrairement au côté client. En plus, il est rarement utilisé pour implémenter des applications basées sur smartphone.

Restrictions de conception : Python est typé dynamiquement, pour cela il n'est pas nécessaire de déclarer le type de variable lors de l'écriture du code.

Couches d'accès aux bases de données sous-développées : Les couches d'accès aux bases de données de Python sont un peu sous-développées par rapport aux autres langages. Pour cela, il est moins souvent appliqué dans les grandes entreprises [64].

4.2.4. Plateforme Anaconda (Distribution Python)

ANACONDA est une distribution libre et open source et un gestionnaire de paquets principalement utilisé dans la science des données avec Python et R et l'apprentissage automatique [61].

Figure IV. 1 Logo d'environnement Anaconda

4.2.4.1. Présentation

Anaconda est une distribution Python dédié à l'analyse de données et calcul scientifique (Par exemple data science, machine Learning applications, large-scale, data processing, predictive analytics). Elle est disponible pour Windows, Mac OS X et Linux, Comprend de nombreux packages populaires : NumPy, SciPy, Matplotlib, Pandas, IPython …En plus elle inclut Spyder, un environnement de développement Python et Conda qui est un gestionnaire de packages indépendant de la plate-forme [65].

4.2.5. Navigateur Anaconda

Il s'agit d'une interface visuelle permettant de lancer les différentes applications de traitement disponibles, de gérer les librairies conda, les environnements et les canaux sans utiliser la moindre ligne de commande. En plus il peut également accéder à des librairies présentes sur le Cloud Anaconda ou dans un Repository Anaconda local, afin de les installer dans un environnement, les exécuter et les mettre à jour. Le navigateur est disponible pour Windows, macOS et Linux [61].

4.2.6. L'invité Anaconda

L'invité Anaconda est une ligne de commande ou un terminal qui permet d'accéder à l'outil conda afin de gérer vos environnements, cette interface vous permet d'installer et de mettre à jour des packages. Il est identique à l'invite de commande sous Windows, la seule différence est qu'ils chargent un environnement pour vous pour plus de commodité. Ceci est similaire à ce que fait Visual Studio, par exemple. Visual Studio installe un lanceur appelé "Invite de commandes Visual Studio" qui fait quelque chose de similaire [66].

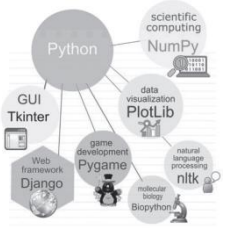
4.2.7. Bases de données utilisées

En informatique, une base de données est un ensemble structuré de données enregistrées sur des supports accessibles par l'ordinateur. Ces informations représentent des données du monde réel et pouvant être interrogées et mises à jour par une communauté d'utilisateurs. Dans notre cas, on a utilisé un fichier de format csv pour stocker notre base de données.

4.3. Librairies et modules

Il existe de nombreuses bibliothèques et modules en Python. Nous pouvons utiliser une bibliothèque ou un module adapté à nos besoins. Par conséquent, les bibliothèques et les modules Python jouent un rôle crucial et sont très utiles aux développeurs.

4.3.1. Qu'est-ce qu'une librairie ou un module python?



Les modules sont des programmes Python (également appelés bibliothèques ou librairies) qui contiennent des fonctions que nous réutilisons fréquemment. Autrement dit, se sont des fichiers script python d'extension .py . Les développeurs Python ont créé de nombreux modules qui effectuent une quantité énorme de tâches. Assurez-vous donc par réflexe qu'une partie du code que vous êtes sur le point d'écrire n'existe pas déjà sous la forme d'un module. La plupart de ces modules sont déjà installés dans la version standard de Python [68].

4.3.1.1. NumPy (Numerical Python)

Est une bibliothèque qui permet d'effectuer des calculs numériques en Python ; elle introduit une gestion facilitée des tableaux de nombres. Est un package de traitement de tableau à usage général et fournit un objet tableau multidimensionnel hautes performances et des outils pour travailler avec ces tableaux. NumPy a résolu le problème de lenteur efficacement. Ces principales caractéristiques sont Fournit des fonctions rapides et précompilées pour les routines numériques, prend en charge une approche orientée objet, le calcul orienté tableau pour une meilleure efficacité et des calculs compacts et plus rapides avec vectorisation. Aussi largement utilisé dans l'analyse de données ainsi que la création d'un tableau puissant à N dimensions. Elle est indispensable pour d'autres bibliothèques telles que scikit-learn et SciPy et lorsqu'il est utilisé avec SciPy et matplotlib remplace le MATLAB [67].

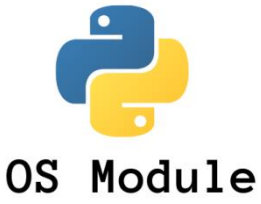
4.3.1.2. Pandas (Panel Data)

Est la bibliothèque la plus populaire conçue pour la manipulation et l'analyse de données en langage Python et utilisé pour la science des données ainsi que pour l'analyse et le nettoyage des données ; elle est incontournable dans le cycle de vie de la science des données. Elle fournit des structures de données flexibles et rapides, et parmi ses principales caractéristiques on trouve qu'elle est une syntaxe éloquente et des fonctionnalités riches qui vous donnent la liberté de gérer les données manquantes, ainsi que sur une série de données vous permettent de créer votre fonction et de l'exécuter, une abstraction de haut niveau et contiennent des structures et des manipulations de données de haut niveau [67].

4.3.1.3. La bibliothèque PIL (Python Imaging Library)

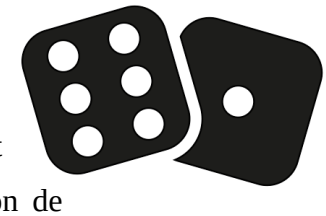
La bibliothèque d'imagerie Python vous permet de créer, modifier et convertir des fichiers image dans une large variété de formats utilisant le langage Python. Cette bibliothèque fournit une prise en charge étendue des formats de fichiers, une représentation interne efficace et des capacités de traitement d'image assez puissantes. La bibliothèque d'images principale est conçue pour un accès rapide aux données stockées dans quelques formats de pixels de base. Il devrait fournir une base solide pour un outil général de traitement d'image [69].

4.3.1.4. Le module OS



Est module fournit un moyen portable d'utiliser les fonctionnalités dépendantes du système d'exploitation. Le module OS de Python fournit des fonctions permettant d'interagir avec le système d'exploitation. Le système d'exploitation fait partie des modules utilitaires standard de Python Les modules « os » et « os.path » incluent de nombreuses fonctions pour interagir avec le système de fichiers [70].

4.3.1.5. Le module Random



Random est un module Python regroupant plusieurs fonctions permettant de générer des nombres aléatoires. Ce sont des nombres pseudo-aléatoires, autrement dit, se sont pas vraiment aléatoires. Ce module peut être utilisé pour effectuer des actions aléatoires telles que la génération de nombres aléatoires, l'impression aléatoire d'une valeur pour une liste ou une chaîne, etc. [71].

4.3.1.6. Le module Tkinter (Tool kit interface)

Est un module intégré à la bibliothèque standard de Python, permettant de créer des interfaces graphiques :

- des fenêtres,
- des widgets (boutons, zones de texte, cases à cocher, …),
- des évènements (clavier, souris, …).

Tkinter est disponible sur Windows et la plupart des systèmes Unix : les interfaces crées avec Tkinter sont donc portables [72].

4.3.1.7. Le module Operator

Est un module disponible dans le cadre de la bibliothèque Python par défaut, fournit des fonctions équivalentes aux opérateurs de Python. Par exemple, vous pouvez multiplier deux nombres en utilisant une fonction au lieu d'utiliser le symbole "*". Ses fonctions sont utiles lorsque vous souhaitez passer des fonctions appelables en tant qu'arguments à un autre objet Python et Certaines d'entre eux peuvent également être utilisées pour effectuer une recherche rapide d'éléments dans des objets de

type itérable.Ainsi que les fonctions d'opérateur fournissent une syntaxe beaucoup plus propre et plus courte [73].

4.4. Les fonctionnalités du système

A ce stade, nous allons décrire l'aspect fonctionnel de l'application développée tout en illustrant par des captures d'écran. La Figure 2 représente l'interface d'accueil de notre application.

Figure IV. 2 l'interface d'accueil.

Interface « Prédire le vainqueur de la coupe du monde actuelle » :

La figure 3 représente une interface qui nous permet d'afficher le vainqueur de la coupe du monde actuelle « Coupe du monde 2022 ».

Figure IV. 3 Interface de Prédiction du vainqueur de la coupe du monde actuelle.

Interface « Prédire l'issue d'un match » :

La figue 4 représente une interface qui nous permet de choisir un tirage fixe ou aléatoire pour le calcul de l'issue de chaque match de la coupe du monde actuelle et le pourcentage de ces derniers pour la gagner.

Figure IV. 4 Interface de Prédiction l'issue d'un match de la coupe du monde actuelle.

Interface « Afficher l'issue d'un match » :

La figure 5 représente une interface qui permet l'affichage du résultat d'un match et le pourcentage de chaque équipe concurrente de gagner la coupe du monde actuelle.

Figure IV. 5 Interface d'Affichage d'issue d'un match de la coupe du monde actuelle.

Interface « Prédire le vainqueur des autres coupes du monde » :

La figure 6 représente une interface qui permet de remplir le fichier csv on introduisant les données de 32 équipes qualifiées pour pouvoir prédire le vainqueur des autres coupes du monde et l'issue de chaque match. (Dans Notre cas on va choisir la coupe du monde 2018 comme exemple).

Figure IV. 6 Interface de Prédiction du vainqueur des autres coupes du monde.

Interface « La saisie des groupes » :

La figure 7 représente une interface nous permet d'introduire les groupes de la coupe du monde pour qu'on puisse calculer les probabilités de la prédiction dans le cas où l'utilisateur choisi le cas Fixes (la coupe du monde 2018).

Figure IV. 7 Interface de la saisie des groupes.

Interface « d'accueil pour les autres coupes du monde » :

La figure 8 représente une interface qui nous permet de choisir entre prédire le vainqueur d'une coupe du monde choisie (La coupe du monde 2018 comme exemple) et prédire l'issue d'un match.

Figure IV. 8 Interface d'accueil pour les autres coupes du monde.

Interface « Affichage du vainqueur de la coupe du monde » :

La figure 9 représente une interface qui permet l'affichage du vainqueur d'une coupe du monde choisie (notre cas la coupe du monde 2018).

Figure IV. 9 Interface d'Affichage du vainqueur des autres coupes du monde.

Interface « Prédire l'issue d'un match » :

La figure 10 représente une interface qui nous permet de choisir un tirage fixe ou aléatoire pour le calcul de l'issue de chaque match de la coupe du monde choisie et le pourcentage de ces derniers pour la gagner.

Figure IV. 10 Interface de Prédiction d'issue d'un match d'une coupe du monde choisie.

Interface « Affichage d'issue d'un match » :

La figure 11 représente une interface qui affiche l'issue d'un match et la probabilité de chaque équipe choisie de gagner la coupe du monde (On a choisi la coupe du monde 2018).

Figure IV. 11 Interface d'Affichage d'issue d'un match d'une coupe du monde choisie.

4.5. Le calcule de performance de notre système :

Dans Machine Learning, nous avons des mesures de performance pour vérifier les performances de notre modèle. Nous avons diverses mesures de performance telles que la matrice de confusion, la précision, le rappel, le score F1.

La figure suivante montre les résultats du notre modèle et un autre modèle sur un site d'internet [76].

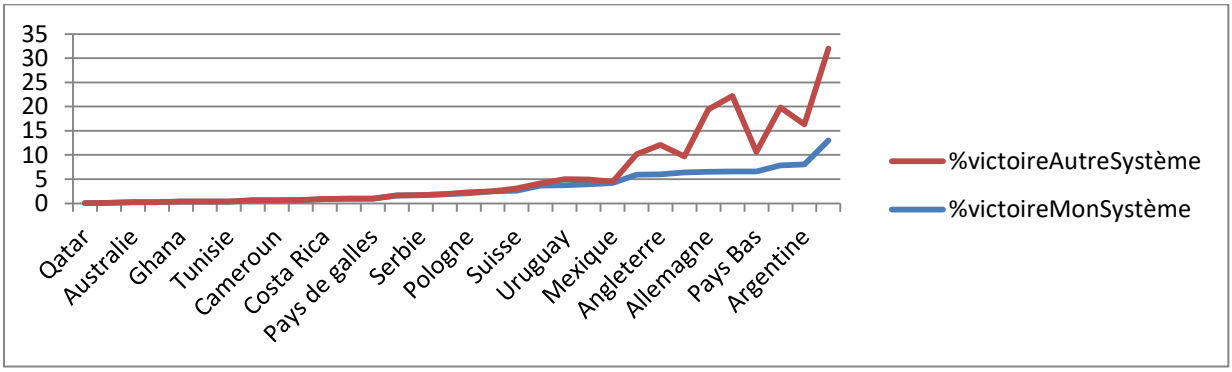


Figure IV. 12 Graphique représentant les résultats probables de notre système et un autre système de prédiction de la coupe du monde 2022.

La matrice de confusion

Une matrice de confusion est une matrice de dimension N * N dans laquelle un axe représente l'étiquette "Réelle" tandis que l'autre axe représente l'étiquette "Prévue". La matrice de confusion est la métrique la plus intuitive et la plus basique à partir de laquelle nous pouvons obtenir diverses autres mesures telles que la précision, le rappel, le score F1.

		Prédiction	
		Positive	**Négative**
Actuelle	**Positive**	**VP**	**FN**
	Négative	**FP**	**VN**

Tableau IV. 1 Matrice de confusion.

Pour une meilleure compréhension de ce que sont VP, FP, VN et FN, nous allons considérer un exemple on veut comparer notre système avec un autre donc :

Négatif : Probabilité de notre système.

Positif : Probabilité de l'autre système.

Vrai positif (VP): Cela représente que l'étiquette prédite est positive et que l'étiquette réelle est positive - correctement prédite. Nous avons prédit que la probabilité de l'autre système. (Positif).

Vrai négatif (VN) : Cela représente que l'étiquette prédite est négative et que l'étiquette réelle est également négative - correctement prédite. Nous avons prédit que la probabilité de notre système. (Négatif).

Faux négatif (FN) : Cela signifie que l'étiquette prédite est négative mais que l'étiquette réelle est positive - prédite à tort. Nous avons prédit que la probabilité de notre système (négatif), et la probabilité de l'autre système (positif).

Faux positif (FP) : Cela signifie que l'étiquette prédite est positive mais que l'étiquette réelle est négative - prédite à tort. Nous avons prédit que la probabilité de l'autre système (positif) mais la probabilité de notre système (négatif) [74,75].

Dans ce qui suit on a une matrice de confusion simplifiée pour mieux calculer les mesures de performances de notre modèle.

pays	VP	FP	VN	FN	presision	recall
Qatar	0	100	0,05	99,95	0	0
Canada	0	100	0,1	99,9	0	0
Australie	0	100	0,22	99,78	0	0
Arabie Saoudite	0	100	0,24	99,76	0	0
Ghana	0	100	0,36	99,64	0	0
Coree du Sud	0	100	0,39	99,61	0	0
Tunisie	0	100	0,4	99,6	0	0
Iran	0	100	0,64	99,36	0	0
Cameroun	0	100	0,66	99,34	0	0
Maroc	0,06	99,94	0,67	99,33	0,0006	0,0006037
Costa Rica	0	100	0,89	99,11	0	0
Ecuateur	0	100	0,96	99,04	0	0
Pays de galles	0	100	0,99	99,01	0	0
Etats unis	0,02	99,98	1,6	98,4	0,0002	0,0002032
Serbie	0	100	1,71	98,29	0	0
Japon	0,02	99,98	1,88	98,12	0,0002	0,0002038
Pologne	0,14	99,86	2,12	97,88	0,0014	0,0014283
Senegal	0	100	2,48	97,52	0	0
Suisse	0,46	99,54	2,65	97,35	0,0046	0,004703
Danemark	0,5	99,5	3,67	96,33	0,005	0,0051637
Uruguay	1,3	98,7	3,73	96,27	0,013	0,0133238
Croatie	1,04	98,96	3,92	96,08	0,0104	0,0107084
Mexique	0,26	99,74	4,23	95,77	0,0026	0,0027075
Portugal	4,22	95,78	5,96	94,04	0,0422	0,0429473
Angleterre	6,1	93,9	5,99	94,01	0,061	0,060933
Belgique	3,28	96,72	6,39	93,61	0,0328	0,0338528
Allemagne	12,88	87,12	6,55	93,45	0,1288	0,1211323
Espagne	15,6	84,4	6,57	93,43	0,156	0,1430799
Pays Bas	4	96	6,61	93,39	0,04	0,041072
France	11,96	88,04	7,87	92,13	0,1196	0,1149006
Argentine	8,24	91,76	8,06	91,94	0,0824	0,0822519
Brazil	19	81	13	87	0,19	0,1792453
Total					0,8908	0,8584604

Tableau IV. 2 Matrice de confusion pour notre système.

Les mesures de performance

La précision : la précision mesure la proportion de l'étiquette positive prédite qui est réellement positive [74,75].

Précision =VP/VP +FP= 0.89 (Voir la matrice de confusion).

La haute précision est liée au faible taux de faux positifs. Nous avons une précision de 0,89 , ce qui est plutôt bon.

Rappel (recall) : le rappel mesure la proportion d'étiquettes positives réelles correctement prédites comme positives [74,75].

Rappel = TP/TP+FN = 0.86 (Voir la matrice de confusion).

Nous avons un rappel de 0,86, ce qui est bon pour ce modèle car il est supérieur à 0,5.

Calcule de F1 Score : Le score F1 est une autre des bonnes mesures de performance qui exploite à la fois les mesures de précision et de rappel. Le score F1 peut être obtenu en prenant simplement la «moyenne harmonique» de la précision et du rappel. Contrairement à la précision qui se concentre principalement sur les faux positifs et au rappel qui se concentre principalement sur les faux négatifs, le score F1 se concentre à la fois sur les faux positifs et les faux négatifs [74,75].

Score F1 = 2*(Rappel * Précision) / (Rappel + Précision)

Score F1 = 2*(0.86* 0.89) / (0.86 + 0.89)

Score F1 = 0,87

Donc à base de rappel et précision, on déduit que notre système est performant.

4.6. Conclusion

Dans ce chapitre nous avons, au premier lieu, présenté les différents outils et langages que nous avons utilisé pour implémenter notre application. Par la suite, nous avons présenté quelques interfaces de notre application et une évaluation de performance de notre système.

Conclusion et perspectives

Conclusion et perspectives

Ce mémoire de fin d'études a eu pour objectif de répondre à la question, « peut-on vraiment prédire le vainqueur de la coupe du monde ? ». Pour conclure on commencera par citer les principales tâches qui nous ramènent au but principal convoité par ce mémoire qui était la conception d'un système de prédiction du vainqueur de la coupe du monde.

Plusieurs algorithmes de machine Learning seront d'une grande utilité pour la prédiction, parmi eux on a utilisé celui de monte Carlo, et le bon processus de cet algorithme se base essentiellement sur le bon classement des données Fifa dans un fichier csv.

Et même si on n'arrive pas à prédire l'issue de tous les matchs avec une précision infaillible en utilisant les algorithmes de machine Learning dans certains cas puisque y a une part de surprise et d'imprévisible conséquente dans les matchs de la coupe du monde, mais on a essayé avec la méthode de monte Carlo d'obtenir des résultats de prédictions du niveau d'un être humain qui suit le football régulièrement et connaît bien la plupart des équipes.

Bref, On a commencé par la récolte des données de la Fifa quelques soit les équipes et leurs classement Fifa ainsi que le nombre de finals jouées et gagnées depuis l'apparition de la coupe du monde dans un fichier en format csv. Puis, on traite ces données à l'aide d'un algorithme de machine Learning « algorithme de monte Carlo » afin de prédire les probabilités de chaque équipe pour gagner la coupe de monde et la probabilité de chacune pour gagner un match.

Et quand on a mis en place le système de prédiction, on a rencontré plusieurs difficultés on présente dans ce qui suit les plus importantes

1- Nous sommes rendus compte que les données Fifa de la coupe du monde 2022 pour remplir notre base de données ne sera pas possible jusqu'à ce que les matchs de barrages terminent ‘ On a terminé le remplissage de la base de données le 14-06-2022‘
2- On n'a pas pu tirer directement nos données manuellement de site officiel de la Fifa puisqu'il y a des cas exceptionnels tels que la coupe de monde 1938 ‘le nombre d'équipes qualifiées c'est 15 pas 32 ‘
3- On a rencontré certaines erreurs lors de l'installation des packages.
4- L'exécution lente de l'application puisque le pc n'est pas performant.

Enfin, les perspectives de nos travaux sont multiples :

- On peut remplacer le fichier csv qui est enregistrée sous forme d'une base de données desktop par une base de données enregistrée sur un serveur et la charger avec un API, ce qui permet de faire une version Android de l'application.
- Stocker l'application sur un serveur web.
- Nous souhaitons rendre notre système de prédiction compatible avec plusieurs plateformes mobiles.

Cette contribution n'est qu'une esquisse qui va être développée ultérieurement. En effet ce travail étant un essai, n'est donc pas un système unique et parfait, c'est pourquoi nous restons ouverts à toutes les critiques et nous sommes prêts à recevoir toutes les suggestions et remarques tendant à améliorer d'avantages cette initiative.

Références

Références :

[1] Gaspard, Amaury TISSEAU, Aline BENABBOU, Rida ENNAMIRI.(2020). RAPPORT TECHNIQUE Paris sportifs Football. École nationale supérieure mines-télécom atlantique Bretagne pays de la Loire.

[2] Despagne, Wilfried. (2010). Construction, analyse et implémentation d'un modèle de prévision. Déploiement sous forme d'un système de prévision chez un opérateur européen du transport et de la logistique. Diss. Université Européenne de Bretagne.

[3] Ali, R. (2022, 18 mars). Predictive Modeling: Types, Benefits, and Algorithms .Oracle NetSuite. Disponible à l'adresse : https://www.netsuite.com/portal/resource/articles/financial-management/predictive-modeling.shtml. Consulté le 05/2022.

[4] Santi, J. S. M. (2018, 26 avril). L'utilisation de la logique de prédiction. Disponible à l'adresse : https://www.e-marketing.fr/Thematique/academie-1078/fiche-outils-10154/L-utilisation-de-la-logique-de-prediction-324698.htm. Consulté le 05/2022.

[5] Solutions Business Intelligence. (2022, 17 juin). L'analyse prédictive : sa définition et ses enjeux pour les entreprises. Disponible à l'adresse : https://solutions-business-intelligence.fr/le-dico-bi/analyse-predictive/. Consulté le 06/2022.

[6] A. (2022, 16 février). Ces prédictions sur la Coupe du monde peuvent ou ne pas se réaliser, mais sont hautement improbables. CONGO CHECK. Disponible à l'adresse : https://congocheck.net/ces-predictions-sur-la-coupe-du-monde-peuvent-ou-ne-pas-se-realiser-mais-sont-hautement-improbables/ . Consulté le 07/2022.

[7] (2022). The Advantages and Disadvantages of Using Football Prediction Services. In : Online Traveling Guide.Disponible à l'adresse : https://www.thetravelingguide.com/recreation/the-advantages-and-disadvantages-of-using-football-prediction-services/. Consulté le 06/2022.

[8] Herbinet, Corentin. (2018).Predicting football results using machine learning techniques. MEng thesis, Imperial College London .

[9] G. Naga Sujini, B. Poornima and D.Gyana Deepika.(2021 , Juin). FIFA OUTCOME PREDICTION USING MACHINE LEARNING TECHNIQUES. INTERNATIONAL JOURNAL OF COMPUTER ENGINEERING AND APPLICATIONS. 7 (15).

[10] van Rijmenam, M. (2022, 10 juillet). The History of Predictive Analytics-Infographic. Disponible à l'adresse : https://datafloq.com/read/history-predictive-analytics-infographic/. Consulté le 06/2022.

[11] Daniel D. Gutierrez. Guide InsideBIGDATA de l'analyse prédictive.TIBCO Spotfire. Disponible à l'adresse : https://www.celge.fr/wp-content/uploads/2015/12/guide-analyse-prédictive.pdf.

[12] (2013). Machine Learning avec R et Prédictions. In : DATASULTING.Disponible à l'adresse : https://www.datasulting.com/articles/machine-learning-avec-r-et-predictions/.Consulté le 06/2022.

[13] Iseli, C., & Sanchez, E. (1993, April). Spyder: A reconfigurable VLIW processor using FPGAs. In [1993] Proceedings IEEE Workshop on FPGAs for Custom Computing Machines (pp. 17-24). IEEE.

[14] Devigne, A. (2022, July 2). Qu'est-ce que les algorithmes prédictifs ? Kobia. Disponible à l'adresse : https://kobia.fr/quest-ce-que-les-algorithmes-predictifs/. Consulté le 06/2022

[15] (1997) TM Mitchell, « Apprentissage automatique WCB » : McGraw-Hill Boston.

[16] ABDOULAYE, A. H., & HOUNDJI, V. R. Détermination des paramètres d'une heuristique générique pour les jeux de type «n-alignés» par apprentissage automatique.

[17] Ouail, B., & Adenane, H. (2021). Etude et comparaison de modèles de prédiction basés sur l'apprentissage automatique.

[18] Maniraguha, Clément. (2019). Fingerprinting de devices IoT à l'aide de l'apprentissage automatique [thèse de doctorat].

[19] Boateng, E. Y., Otoo, J., & Abaye, D. A. (2020). Basic tenets of classification algorithms K-nearest-neighbor, support vector machine, random forest and neural network: a review. Journal of Data Analysis and Information Processing, 8(4), 341-357.

[20] Cortes, C. and Vapnik, V. (1995) Support-Vector Networks. Machine Learning, 20, 273-297.

[21] Abedi, M., Norouzi, G.H. and Bahroudi, A. (2012) Support Vector Machine for Multi-Classification of Mineral Prospectivity Areas. Computers & Geosciences, 46, 272-283.

[22] Sluiter, R. and Pebesma, E.J. (2010) Comparing Techniques for Vegetation Classification Using Multi- and Hyperspectral Images and Ancillary Environmental Data. International Journal of Remote Sensing, 31, 6143-6161.

[23] Huang, C., Davis, L.S. and Townshend, J.R.G. (2002) An Assessment of Support Vector Machines for Land Cover Classification. International Journal of Remote Sensing, 23, 725-749.

[24] Dietrich, R., Opper, M. and Sompolinsky, H. (1999) Statistical Mechanics of Support Vector Networks. Physical Review Letters, 82, 2975.

[25] Shen, T., Li, H.S., Qian, Z. and Huang, X.L. (2009) Active Volume Models for 3D Medical Image Segmentation. IEEE Conference on Computer Vision and Pattern Recognition, Miami, 20-25 June 2009, 707-714.

[26] Tsai, C.F., Hsu, Y.F., Lin, C.Y. and Lin, W.Y. (2009) Intrusion Detection by Machine Learning: A Review. Expert Systems with Applications, 36, 11994-12000.

[27] Brown, W.M., Gedeon, T.D., Groves, D.I. and Barnes, R.G. (2000) Artificial Neural Networks: A New Method for Mineral Prospectivity Mapping. Australian Journal of Earth Sciences, 47, 757-770.

[28] Porwal, A., Carranza, E.J.M. and Hale, M. (2004) A Hybrid Neuro-Fuzzy Model for Mineral Potential Mapping. Mathematical Geology, 36, 803-826.

[29] Breiman, L. (2001) Random Forests. Machine Learning, 45, 5-32.

[30] Rodriguez-Galiano, V.F. and Chica-Rivas, M. (2014) Evaluation of Different Machine Learning Methods for Land Cover Mapping of a Mediterranean Area Using Multi-Seasonal Landsat Images and Digital Terrain Models. International Journal of Digital Earth, 7, 492-509.

[31] Waske, B. and Braun, M. (2009) Classifier Ensembles for Land Cover Mapping Using Multitemporal SAR Imagery. ISPRS Journal of Photogrammetry and Remote Sensing, 64, 450-457.

[32] Chen, W., Xie, X.S., Wang, J.L., Pradhan, B., Hong, H.Y., Bui, D.T., Duan, Z. and Ma, J.Q. (2017) A Comparative Study of Logistic Model tree, Random Forest, and Classification and Regression Tree Models for Spatial Prediction of Landslide Susceptibility. Catena, 151, 147-160.

[33] Statnikov, A., Wang, L. and Aliferis, C.F. (2008) A Comprehensive Comparison of Random Forests and Support Vector Machines for Microarray-Based Cancer Classification. BMC Bioinformatics, 9, Article No. 319.

[34] Hmeidi, I., Hawashin, B. and El-Qawasmeh, E. (2008) Performance of KNN and SVM Classifiers on Full Word Arabic Articles. Advanced Engineering Informatics, 22, 106-111.

[35] Pan, F., Wang, B.Y., Hu, X. and Perrizo, W. (2004) Comprehensive Vertical Sample-Based KNN/LSVM Classification for Gene Expression Analysis. Journal of Biomedical Informatics, 37, 240-248.

[36] Halvani, O., Steinebach, M. and Zimmermann, R. (2013) Authorship Verification via K-Nearest Neighbor Estimation. Notebook PAN at CLEF. CLEF 2013 Working Notes, Valencia, 23-26 September 2013.

[37] Chen, H.L., Huang, C.C., Yu, X.G., Xu, X., Sun, X., Wang, G. and Wang, S.J. (2013) An Efficient Diagnosis System for Detection of Parkinson's Disease Using Fuzzy K-Nearest Neighbor Approach. Expert Systems with Applications, 40, 263-271.

[38] Chan, J.C.W. and Paelinckx, D. (2008) Evaluation of Random Forest and Adaboost Tree-Based Ensemble Classification and Spectral Band Selection for Ecotope Mapping Using Airborne Hyperspectral Imagery. Remote Sensing of Environment, 112, 2999-3011.

[39] Shapire, R.E. and Singer, Y. (1998) BoosTexter: A System for Multi-Label Text Categorization. Machine Learning, 39, 135-168.

[40] Breiman, L. (1996) Bagging Predictors. Machine Learning, 24, 123-140.

[41] Edwin, R. and Bogdan, Z. (2017) Comparison of Support Vector Machine, Random Forest and Neural Network Classifiers for Tree Species Classification on Airborne Hyperspectral APEX Images. European Journal of Remote Sensing, 50, 144-154.

[42] Coimbra, R., Rodriguez-Galiano, V., Olóriz, F. and Chica-Olmo, M. (2014) Regression Trees for Modeling Geochemical Data-An Application to Late Jurassic Carbonates (Ammonitico Rosso). Computers & Geosciences, 73, 198-207.

[43] Wang, Z.L., Lai, C.G., Chen, X.H., Yang, B., Zhao, S.W. and Bai, X.Y. (2015) Flood Hazard Risk Assessment Model Based on Random Forest. Journal of Hydrology, 527, 1130-1141.

[44] Sun, L. and Schulz, K. (2015) The Improvement of Land Cover Classification by Thermal Remote Sensing. Remote sensing, 7, 8368-8390.

[45] Vincenzi, S., Zucchetta, M., Franzoi, P., Pellizzato, M., Pranovi, F., De Leo, G.A. and Torricelli, P. (2011) Application of a Random Forest Algorithm to Predict Spatial Distribution of the Potential Yield of Ruditapes philippinarum in the Venice Lagoon, Italy. Ecological Modelling, 222, 1471-1478.

[46] Brédy, J. (2019). Prévision de la profondeur de la nappe phréatique d'un champ de canneberges à l'aide de deux approches de modélisation des arbres de décision.

[47] Ange-Boris BRIKA.(2014). Data Mining avec Weka. Ecole Polytechnique de Montréal.

[48] Azzaz, S., & Namous, R. (2022). Étude comparative des algorithmes d'apprentissage de la machine dans la prédiction et diagnostic du cancer de sein (Doctoral dissertation, Université Larbi Tébessi-Tébessa).

[49] linedata(2022). Principaux algorithmes de classification – Partie 1.Disponible à l'adresse : https://fr.linedata.com/principaux-algorithmes-de-classification-partie-1. Consulté le 06/2022.

[50] Help Center. Comparaison d'algorithmes de Machine Learning supervisé.Disponible à l'adresse : https://help.xlstat.com/fr/6517-comparison-supervised-machine-learning-algorithms. Consulté le 06/2022.

[51] Setiono R. and Loew, W. K. (2000), FERNN: An algorithm for fast extraction of rules from neural networks, Applied Intelligence.

[52] Witten, I. H. & Frank, E. (2005). Data Mining: Practical machine learning tools and techniques (2nd ed.), ISBN: 0- 12-088407-0, Morgan Kaufmann Publishers, San Francisco, CA, U.S.A. © 2005 Elsevier Inc.Retrieved from website:

ftp://93.63.40.27/pub/manuela.sbarra/Data Mining Practical Machine Learning Tools and Techniques - WEKA.pdf

[53] Sidahmed Amel, Rabhi Karima. (2020). La prédiction du diabète en utilisant les algorithmes de machine Learning [Mémoire non publié].

[54] EZZIKOURI, H., & Fakir, M. Algorithmes de classification: ID3 et C4. 5. Url: https://www.academia.edu/33701469/AlgorithmesdeclassificationID3andC, 4,5.

[55] Salaht, F. A., and Djabali, Y. (2008).Stablilité forte, simulation et performance de la méthode dans un système de files d'attente M/G/1 avec service par groupe à capacité aléatoire muni de la politique (r, N) [Mémoire non publié].Université A.Mira de Béjaia.

[56] BAISSA, Kenza. (2012). Simulation Statistique [Thèse de doctorat].

[57] WILL KENTON ,(2021, 4 October). Investopedia. In: What Is a Monte Carlo Simulation .Disponible à l'adresse: https://www.investopedia.com/terms/m/montecarlosimulation.asp. Consulté le 05/2022.

[58] LYNDA Bouhidel.(juin 2012). La méthode monte Carlo pour l'analyse d'un système de production (aspect dysfonctionnel) [thèse de magister]. université El-Hadj Lakhdar-Batna,

[59] MEHIRA, C. (2021). Influence de l'Enveloppe architecturale sur la Performance Énergétique des bâtiments.

[60] Van Rossum, G. (2003). An introduction to Python (p. 115). F. L. Drake (Ed.). Bristol: Network Theory Ltd.

[61] TERGHINI, H. Conception et réalisation d'une approche de Deep Learning pour l'IoT, Université Mohamed Khider – BISKRA ,2020.

[62] L, B. (2022, 11 juillet). Python : tout savoir sur le principal langage Big Data et Machine Learning .LeBigData.fr. Disponible à l'adresse : https://www.lebigdata.fr/python-langage-definition#:~:text=Le%20principal%20cas%20d%27usage,seules%20utilités%20de%20ce%20langage. Consulté le 06/2022.

[63] Derfoufi, Y. (2019). Formation en langage Python.Disponible à l'adresse : https://www.tresfacile.net/doc/python/formation/part1/Formation-Python-Chapitre1.pdf.

[64] Khoirom, S., Sonia, M., Laikhuram, B., Laishram, J., & Singh, T. D. (2020). Comparative analysis of Python and Java for beginners. Int. Res. J. Eng. Technol, 7(8), 4384-4407.

[65] Rolon-Mérette, D., Ross, M., Rolon-Mérette, T., & Church, K. (2016). Introduction to Anaconda and Python: Installation and setup. Python for research in psychology, 16(5), S5-S11.

[66] Difference between Anaconda Prompt, Command Prompt, Git Shell Etc.. Disponible à l'adresse :https://www.reddit.com/r/learnprogramming/comments/cc98y3/difference_between_anaconda_prompt_command_prompt/. Consulté le 06/2022.

[67] Sayeth Saabith AL, Vinothraj T, Fareez MMM.(2020,Novembre).POPULAR PYTHON LIBRARIES AND THEIR APPLICATION DOMAINS. International Journal of Advance Engineering and Research Development.7(11).

[68] Patrick Fuchs, Pierre Poulain, version. (2016). Cours de Python. univ-paris Diderot.

[69] John W. Shipman.(2013). Python Imaging Library (PIL). New Mexico Tech Computer Center.

[70] GeeksforGeeks. (2022, June 16). OS Module in Python with Examples. Disponible à l'adresse : https://www.geeksforgeeks.org/os-module-python-examples/. Consulté le 07/2022.

[71] GeeksforGeeks. (2021, December 14). Python Random Module.Disponible à l'adresse : https://www.geeksforgeeks.org/python-random-module/. Consulté le 07/2022.

[72] Site de Patrick Darcheville (*juillet 2022*) .Le module Tkinter pour réaliser de superbes interfaces graphiques.Disponible à l'adresse : https://darchevillepatrick.info/python/python19.php. Consulté le 06/2022

[73] Linuxhint (2021). How to Use Operator Module in Python. . Disponible à l'adresse : https://linuxhint.cossm/use-operator-module-python/. Consulté le 06/2022.

[74] Precision, Recall et Precision-Recall curve. (2 Juillet 2022). Kobia.Disponible à l'adresse : https://kobia.fr/classification-metrics-precision-recall/. Consulté le 09/2022.

[75] Tremblay, C. (2 Juillet 2022). F1-score, la synthèse entre precision et recall. Kobi .Disponible à l'adresse : https://kobia.fr/classification-metrics-f1-score/. Consulté le 09/2022.

[76] Raghunathan, A. (31 Decembre 2020,). Simulating the FIFA World Cup 2022 .Par Abhinav Raghunathan | Towards Data Science. Medium.Disponible à l'adresse : https://towardsdatascience.com/simulating-the-fifa-world-cup-2022-d363fad7da22. Consulté le 09/2022.

Résumé

Des systèmes prédictifs ont été utilisés pour prévoir des événements et des résultats pratiquement dans tous les domaines de la vie. La prédiction des résultats de football en particulier a gagné en popularité ces dernières années. Dans notre mémoire, On va faire une prédiction à l'aide d'un algorithme basait cependant exclusivement sur les classements Fifa de chaque équipe. En plus de cela, nous avons opté pour un modèle de Machine Learning supervisé de Naïve bayes pour la prédiction. Notre système de prédiction des résultats des matchs de football de la coupe de monde a été mis en œuvre à l'aide d'un algorithme probabiliste qui est l'algorithme de Monte Carlo.

Les mots clés : Système prédictif, Machine Learning supervisé, Monte Carlo.

Abstract

Predictive systems have been used to predict events and results in every area of life. The prediction of football results in particular has popularity in recent years. In our final dissertation, we will make a prediction using an algorithm based on the Fifa rankings of each team. In addition to this, we opted for a supervised Machine Learning model of Naïve Bayes for the prediction. Our system for predicting the results of world cup football matches has been implemented using a probabilistic algorithm which is the Monte Carlo algorithm.

Key words: Predictive system, Supervised Machine Learning, Monte Carlo.