République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
École supérieure en sciences et technologies de l'informatique et du numérique

Rapport du projet Apprentissage Machine

Préparé par

M^{lle} **ALKAMA Narima**

M^{lle} **BOUROUBA Asma**

M^{lle} **HAMDANE Meriem**

Startup success prediction

Encadrante:

M^{me}DAOUDI Meroua

Année Universitaire: 2022/2023

Table des matières

Table des figures

1 Introduction

Le machine learning, avec son pouvoir prédictif, ouvre de nouvelles perspectives passionnantes dans de nombreux domaines. Grâce à l'analyse approfondie des données et à la détection de modèles complexes, le machine learning permet de créer des modèles prédictifs précis. Ces modèles peuvent être utilisés pour anticiper et prédire des événements futurs, des comportements des utilisateurs, des tendances du marché, et bien plus encore. Que ce soit dans le domaine des sciences, de la finance, de la santé ou du commerce, le machine learning apporte des outils puissants pour la prise de décision éclairée. En exploitant les vastes quantités de données disponibles, le machine learning transforme les informations en connaissances exploitables, offrant ainsi un avantage concurrentiel et ouvrant de nouvelles opportunités. Les prédictions du machine learning peuvent aider à optimiser les processus, à améliorer les produits et services, et à prendre des décisions stratégiques basées sur des analyses précises. Grâce à son pouvoir prédictif, le machine learning révolutionne notre capacité à comprendre et à façonner le monde qui nous entoure.

2 Définition du problème

Une startup est une entreprise ou un projet lancé par un entrepreneur pour rechercher, développer et valider un modèle économique évolutif. Les startups sont confrontées à une grande incertitude et ont un taux d'échec élevé, mais une minorité d'entre elles réussissent et deviennent influentes.

Les startups ont connu une croissance exponentielle au cours des dernières années. Prédire le succès d'une startup permet aux investisseurs de trouver des entreprises qui ont un potentiel de croissance rapide, ce qui leur permet d'avoir une longueur d'avance sur la concurrence.

L'objectif de notre projet est d'analyser le comportement des startups en fonction de plusieurs variables, de déterminer quelles variables affectent le plus le succès des startups, puis de construire un modèle capable de prédire le succès d'une startup.

3 Explication de la dataset

Afin d'atteindre l'objectif de notre projet nous allons utiliser les techniques d'apprentissage automatique pour analyser les caractéristiques ou les variables associées aux startups que nous

avons collecté grâce au jeu de données (dataset) à partir de la plateforme Kaggle. Ce dataset est de 49 colonnes et 923 lignes et certaines de ses variables sont les suivantes :

- **Latitude :** la latitude d'une startup est une mesure de son potentiel de croissance et de succès, et est souvent une considération clé pour les investisseurs et les parties prenantes lors de l'évaluation des perspectives de la startup.
- **Longitude :** c'est la position de la startup sur le marché ou l'industrie par rapport à ses concurrents. Les startups avec une longitude élevée ont moins de concurrence sur le marché. et ceux dont la longitude est faible peuvent faire face à une concurrence intense et avoir du mal à se différencier des concurrents.
- **Labels :** celui-ci a les valeurs 0 ou 1, 1 pour une startup acquise et 0 pour celui qui a échoué.
- **founded_at :** date de fondement de la startup.
- **closed_at :** date de fermeture.
- **first_funding_at :** la date de réception de son premier aide financier généralement fournis par des investisseurs providentiels, des amis et des membres de la famille ou early-stage venture capital firms.
- **Last_funding_at :**la dernière fois qu'il reçoit un financement.
- **age_first_funding_year :**est l'âge de l'entreprise en années depuis qu'il a obtenu son premier financement.
- **age_last_funding_year :** est l'âge de l'entreprise en années depuis qu'il a obtenu son dernier financement
- **age_first_milestone_year :** Une étape importante pour une startup est une réalisation ou un événement significatif qui marque la progression vers les objectifs de l'entreprise tels que : Lancer un produit minimum viable (MVP), Sécuriser le financement (la levée de fonds auprès d'investisseurs peut être une étape majeure pour une startup), Acquérir des clients, Atteindre Rentabilité et expansion : développer l'équipe, s'étendre sur de nouveaux marchés ou lancer de nouveaux produits ou services. l'âge signifie l'âge de la startup lorsqu'elle atteint son premier jalon.
- **relationships :** il indique combien de relations une startup a-t-elle. Par exemple, une start-up peut avoir des relations avec des comptables, des investisseurs, des fournisseurs, des mentors, etc.

- **funding_rounds :** pour une startup est une période de temps pendant laquelle l'entreprise cherche à lever des capitaux auprès d'investisseurs en échange de capitaux propres dans l'entreprise.
- **milestones :** nombre total de milestones que la startup a au cours de sa vie.
- **state_code :** contient le code de l'état.
- **is_CA/ is_NY / is_MA / is_TX . . . :** des champs binaires qui montrent si la startup appartient ou pas à cet état.
- **category_code :** code de la category.
- **is_software / is_mobile . . .. :** colonnes binaires pour les catégories.
- **is_VC / is_angel . . . :** se sont les types de fonds.
- **is_VC :** VC signifie capital-risque, qui est un type d'investissement en capital-investissement effectué dans des entreprises en démarrage qui ont un potentiel de croissance et de rendement élevés. Les capital-risqueurs sont généralement des investisseurs professionnels qui gèrent des fonds dédiés à l'investissement dans des startups et d'autres entreprises à forte croissance.
- **is_angel :** sont généralement des personnes fortunées qui fournissent un financement de démarrage aux startups en échange d'actions dans l'entreprise. Ils investissent souvent dans des startups qui en sont aux tout premiers stades de développement, tels que le stade de pré-amorçage ou d'amorçage, et cherchent à lever des capitaux pour lancer ou développer leur produit ou service.
- **is_roundA :** c'est une étape importante car elle fournit le capital nécessaire pour faire évoluer l'entreprise et progresser vers la rentabilité. Il valide également le business model de la startup et apporte à l'entreprise le soutien et l'expertise de ses investisseurs.
- **is_roundB :**est le prochain cycle de financement qu'une startup reçoit après son cycle de financement de série A. À ce stade, la startup a généralement atteint un certain niveau de succès et recherche des capitaux supplémentaires pour faire évoluer ses opérations et accroître sa part de marché.
- **is_roundC :** est le prochain cycle de financement qu'une startup reçoit après son cycle de financement de série B. À ce stade, la startup a généralement atteint une croissance et une traction sur le marché importantes et recherche des capitaux supplémentaires pour étendre davantage ses opérations, ses offres de produits et sa part de marché.

- **is_roundD :** est le prochain cycle de financement qu'une startup reçoit après son cycle de financement de série C. À ce stade, la startup a généralement atteint une croissance et une domination du marché significatives et recherche des capitaux supplémentaires pour maintenir son avantage concurrentiel, se développer à l'échelle mondiale ou explorer de nouveaux secteurs d'activité.
- **is_top500 :** valeur binaire 1 si la startup est parmi les top 500 , 0 sinon.
- **Status :** la target, 1 si la startup est une réussite , 0 sinon.

4 Outils et packages utilisés

Dans le but de mise en œuvre d'un modèle d'apprentissage machine pour la prédiction du succès d'une startup nous avons utilisé le langage Python avec la plateforme de développement et d'exécution de code Google Colab afin de faciliter la collaboration entre les membres de l'équipe.

Le choix du langage de programmation s'est porté sur Python dû aux différents packages qu'il offre, parmi ceux la nous citons :

1. NumPy : NumPy est une bibliothèque fondamentale pour le calcul scientifique en Python. Elle offre des structures de données de tableaux multidimensionnels performantes et des fonctions mathématiques pour effectuer des opérations sur ces tableaux.
2. Pandas : Pandas est une bibliothèque utilisée pour la manipulation et l'analyse de données. Elle fournit des structures de données flexibles et performantes, comme les DataFrames, qui facilitent le chargement, la manipulation et la transformation de données tabulaires. Nous l'avons utilisé dans l'analyse de données et la préparation des données pour l'apprentissage automatique.
3. Matplotlib : Matplotlib est utilisée pour la visualisation des données en 2D. De même pour Seaborn, elle offre une interface de haut niveau pour créer des graphiques statistiques attrayants et informatifs. Nous l'avons utilisé pour explorer et visualiser des relations complexes entre plusieurs variables du dataset.
4. Scikit-learn : Scikit-learn est une bibliothèque très utile avec une large gamme d'algorithmes et d'outils pour le prétraitement des données, la réduction de dimension, la classification, la régression, le regroupement, etc.

5 Plan suivi pour la construction du modèle

5.1 Fixation d'un objectif mesurable

Réalisation d'une prédiction avec une bonne métrique de performance, dans notre cas nous avons fixé la valeur du recall à 90%.

5.2 Analyse et exploration des données

L'Analyse et l'exploration des données est une étape cruciale afin de se familiariser avec le dataset et comprendre au maximum les différentes variables pour définir la stratégie de modélisation.

Ce processus comprend deux étapes :

5.2.1 Analyse de la forme

Cette étape comprend les opérations suivantes :

1. Identification de la variable cible (Target variable) : la variable status est la variable cible, une variable catégorielle qui prend les valeurs "acquired" ou "closed".
2. Nombre de lignes et de colonnes du dataset : Ce dataset contient 49 colonnes, et 923 lignes non dupliquées.
3. Types de variables : Concernant le type des variables, nous trouvons 28 caractéristiques de type entier, 14 variables catégorielles et 7 de type float.
4. Identification des valeurs manquantes : En effectuant des calculs et des visualisations sur le dataset nous avons constaté que 5 colonnes contiennent des valeurs manquantes.

5.2.2 Analyse du fond

L'étape de l'analyse du fond résume les opérations suivantes :

1. visualisation de la variable target : La visualisation de la variable cible "status" nous donne 65% des startups sont acquired alors que 35% sont closed
2. Compréhension des différentes variables en se documentant sur internet :
   - variables quantitatives non standardisées à l'exception des colonnes age_milstone_first_fundings et age_milstone_last_fundings.
   - variables qualitatives sont multicatégories.

3. Visualisation des relations variables/target

   - Status/Latitude : on constate que le nombre de startups acquis avec une grande latitude est un peu plus élevé que les latitudes des startups fermées. On peut dire que la latitude est une bonne métrique pour prendre des décisions qui satisfont notre objectif, mais pas tous le temps.
   - Status/State_code : on remarque que pour CA qu'il existe un très grand nombre de startup dans cet état. En plus, une startup à plus de chance pour qu'elle survit par rapport à d'autre états. Les startups de NY ont aussi plus de chance de survivre, malgré le nombre considérable de startup fermés. Les résultats observés dans NY , CA et MA montre qu'il y a une concurrence dans le marché de ces états.
     En conclusion, cette variables semblent être importantes pour notre modèle
   - Status /Category : Les startups de catégorie : web - mobile - software et entreprise sont les plus vivante sur le marché.
   - Status/Autre variables :
     - labels : on ne sait pas que signifie cette colonne mais d'après ce que montre la heatmap, le label est donné pour les startups acquis. cette colonne représente exactement les même signification que la colonne status donc on en aura pas besoin.
     - funding_rounds : on remarque que ce critère n'est pas fiable car les startup qui appartiennent aux deux classes (fermées - acquits) montre presque la même corrélation avec la target status
     - milestones : on remarque que la majorité des startups acquises avaient entre 1 et 3 milestones. cela peut nous être intéressant
     - is_CA : CA contient un très grand nombre de startups dont la majorité sont des startups acquises, ensuite NY après MA. Cette colonne peut nous être utile.
     - le type de fond : les startups qui ont eux des fonds du type roundA ont réussi à survivre plus longtemps que celles qui ont bénéficié d'autres types de fonds.
     - is_top500 : la majorité des startups qui ont été classées top 500 sont des startups acquises.

4. Visualisation des relations variables/variables : les colonne labels, funding_rounds sont des colonnes inutiles à notre objectif
5. Identification des outliers : A partir de la description détaillée du dataset nous remarquons que les colonnes longitude, age_first_funding_year, age_last_funding_year, age_first_milestone_year et age_last_milestone_year comprennent des valeurs négatives.

5.3 Pré-traitement des données

Cette étape vise à préparer les données en effectuant des transformations et des ajustements pour garantir la qualité des données et les rendre adaptées à l'analyse.

Lors de ce processus nous allons passer par différentes techniques pour mettre le dataset dans un format propice au développement du modèle de Machine Learning.

5.3.1 Traitement des valeurs manquantes

L'ensemble des valeurs manquantes ont été traités comme suit :
Basé sur les résultats obtenus de l'analyse , les colonnes 'age_first_milestone_year' et 'age_last_milestone_year' ont des valeurs nulles lorsque la startup n'a pas de milstones, ceci peut être confirmé en regardant la colonne 'milestones' contenant les valeurs 0 dans les lignes contenant NaN dans les colonnes 'age_first_milestone_year' et 'age_last_milestone_year'. Nous avons donc décidé de remplir les valeurs manquantes avec la valeur 0.

5.3.2 Traitement des valeurs aberrantes

A partir de la description détaillée du dataset nous remarquons que les colonnes longitude, age_first_funding_year, age_last_funding_year, age_first_milestone_year et age_last_milestone_year comprennent des valeurs négatives.

nous avons donc décidé de convertir les valeurs négatives en des valeurs null (zéros), vu que les tuples contennent des valeurs négatives sur la colonne age_first_funding_year ont recus leurs premier fond avant leurs création. Pareil pour les autres colonnes, les valeurs négatives ont été converties en des valeurs null (zéros).

5.3.3 Suppression des caractéristiques non nécessaires

- **State_code.1 :** la colonne "state_code" et la colonne "state_code.1" sont identiques, donc la colonne "state_code.1" doit être supprimée.
- **Unnamed :6 :** La colonne "Unnamed : 6" est une combinaison des colonnes "city", "state_code", and "zip_code". Nous avons donc décidé de supprimer cette colonne.
- **closed_at :** cette colonne contient plus de 60% de valeurs manquantes.L'utilité de son imputation nous a été floue ,en plus du fait que dans le cas général la date de fermeture des startups ne représente pas un critère fort pour notre étude.Pour cela, on a décidé de supprimer cette colonne. c'est pareil pour la colonne founded_at.
- **Longitude :** cette interprétation de la longitude pour les startups n'est pas un concept largement utilisé et n'est pas une mesure standard du potentiel ou du succès d'une startup.
- **labels :** le label est donné pour les startups acquis. Cette colonne représente exactement les mêmes significations que la colonne status donc on n'en aura pas besoin.
- **funding_rounds :** on remarque que ce critère n'est pas fiable car les startup qui appartiennent aux deux classes (fermées - acquis) montrent presque la même corrélation avec la target status.
- **id, object_id :** ces colonnes n'apportent pas d'information à notre modèle.
- **category_code , state_code , zip_code , city :** sont des colonnes encodées par la méthode de l'encodage binaire, ce qui veut dire que les valeurs de ces colonnes existent comme caractéristiques (colonnes features) donc on aura pas besoins de ces quatres colonnes.

5.3.4 Encodage des variables catégoriques

Les variables catégoriques de notre dataset : category_code , city , state_code ainsi que les types de fonds sont déjà encodés.

Après que nos données soient prêtes à la modélisation. On cherche à bien choisir un modèle qui agit de la bonne manière sur nos données.Pour cela, la prochaine étape est la sélection du modèle.

5.4 Sélection du modèle

5.4.1 Phase de sélection

Cette phase de construction du modèle consiste à sélectionner parmi plusieurs modèles, celui qui agit bien sur nos données. Après l'évaluation de plusieurs modèles grâce à la fonction cross validation que montre la figure 1, nous avons réussi à sélectionner 3 modèles qui possèdent les meilleurs scores sur nos données, qui sont : Random Forest Classifier, Gradient Booster Classifier et Decision Tree Classifier. Puis nous avons entraîner ces modèles sur nos données pour voir quelle modèle donnera le meilleur testing accuracy, pour qu'au final nous avons eu les résultats suivants :

Testing accuracy for :

- Random Forest Classifier : 77
- Gradient Boosting Classifier : 74
- Decision Tree Classifier : 67

C'est d'après ces résultats qu'on a opté d'utiliser Random Forest Classifier comme modèle pour notre projet.

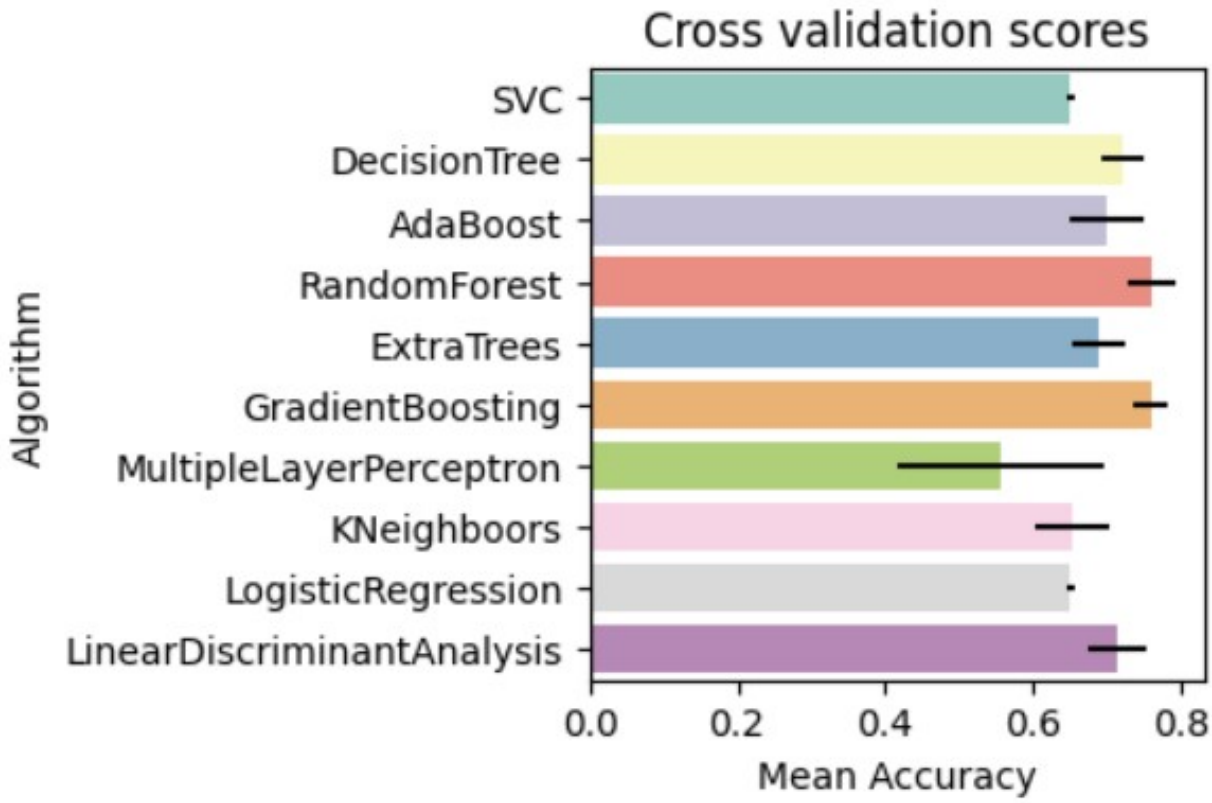

FIGURE 1 – Cross validation scores

5.4.2 Définition du modèle

Un random forest classifier (classifieur de forêt aléatoire) est un algorithme d'apprentissage automatique utilisé pour la classification et la régression. Il fait partie de la famille des méthodes d'ensemble, ce qui signifie qu'il combine les prédictions de plusieurs modèles individuels pour produire une prédiction finale.

5.4.3 Feature selection

Après le choix du modèle on a essayé d'améliorer la performance du modèle en sélectionnant un échantillon de caractéristiques par les méthodes de réduction de dimensionnalité citant : RFE (Recursive feature elimination) , PCA (Principal Component Analysis).

Ces méthodes là ainsi que les données propres (sans sélection ou réduction de dimension) ont été tester sur le modèle RFC (Random Forest Classifier) et été évalués par la validation croisée ce qui a donner les résultats suivant :

- **RFC sans sélection** :
  - Accuracy : 79.06
  - Cross-validation scores : [0.74615385 0.79844961 0.76744186 0.75968992 0.72093023]
  - Cross-validation mean : 75.85
- **RFE (Recursive Feature Elimination) :**
  - Accuracy : 78.34
  - Cross-validation scores : [0.73076923 0.79069767 0.75193798 0.76744186 0.75968992]
  - Cross-validation mean : 76.01
- **PCA (Principal Component Analysis) :**
  - Accuracy : 70.81
  - Cross-validation scores : [0.71621622 0.68918919 0.66891892 0.65306122 0.6462585]
  - Cross-validation mean : 67.47

C'est sur ces résultats qu'on a décidé de travailler avec toutes les caractéristiques au lieu de sélectionner un échantillon, vu que ça donne de meilleures performances avec la totalité de la dataset.

5.5 Entraînement du modèle

Dans l'entraînement du modèle, on a préféré passer par 4 itérations afin d'atteindre les meilleures performances. dans ce qui suit une explication détaillée de chaque itération :

5.5.1 Itération 1

Dans la première itération, on a choisi de travailler sur toutes les caractéristiques dans notre jeu de données avec le paramètre n_estimators = 100, une valeur choisie après la visualisation d'une courbe montrant l'accuracy de notre modèle en fonction du paramètre n_estimators de Random Forest . Mais les résultats n'étaient pas satisfaisants, donc on a passé directement à la deuxième itération afin d'améliorer ces derniers.

5.5.2 Itération 2

Dans cette deuxième itération, on a remarqué que la qualité de nos données influence directement les performances de notre modèle, et pour régler ce problème, on a décidé d'utiliser la méthode de l'ACP pour visualiser les charges des caractéristiques pour garder seulement celles qui ont des charges élevées.

On a entraîné encore une fois notre modèle avec les mêmes paramètres. Les performances augmentent mais elles ne sont toujours pas satisfaisantes. Donc on passe à une troisième itération.

5.5.3 Itération 3

Dans la troisième itération, on a utilisé la méthode de GridSearch pour chercher les meilleurs paramètres vu que mm avec la réduction de dimension qu'on a faite, les résultats ne nous satisfont pas. Donc cette méthode nous donne les paramètres suivants :
Best parameters : 'max_ depth' : 10, 'n_ estimators' : 300

On les a utilisés dans l'entraînement. Les résultats s'améliorent mais on croit toujours qu'on peut faire mieux. Donc on passe à l'itération suivante

5.5.4 Itération 4

Dans cette dernière itération, on a décidé de réutiliser feature selection avec la méthode de RFE pour sélectionner les 20 caractéristiques les plus signifiantes afin d'améliorer nos résultats au maximum.

5.6 Évaluation du modèle

Pour évaluer notre modèle, on utilise les métriques suivantes :

- **Recall :** est la proportion des objets pertinents proposés parmi l'ensemble des objets pertinents, il est défini par le ratio du nombre de vrais positifs par rapport au nombre total d'objets réels (pertinents).
- **Accuracy :** une métrique de score élémentaire qui calcule le nombre moyen d'observations correctement prédites
- **La courbe ROC :** signifie l'ajustement entre le TPR (taux positif vrai) et le FRP (taux positif faux)
- **La matrice de confusion :** une matrice qui résume les résultats de prédiction de notre problème particulier de classification. Elle compare les données réelles pour une variable cible à celles prédites par un modèle.

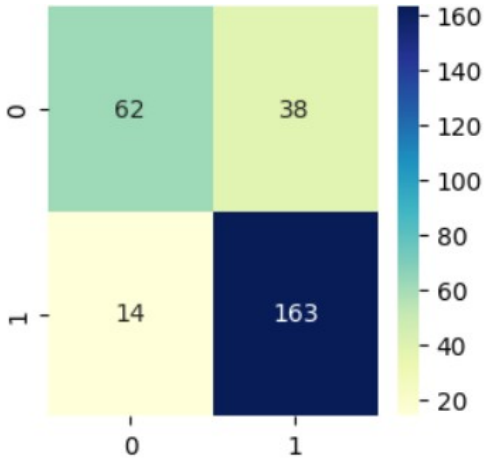

FIGURE 2 – Matrice de confusion

5.7 Réglage du modèle

Pour améliorer notre modèle, on avait besoin de régler ses paramètres. Pour celà, on utilise les méthodes suivantes :

- Utiliser une courbe d'accuracy du modèle en fonction du paramètre n_estimators et choisir la valeur où l'accuracy est au maximum comme le montre la figure 3.
- Sélection des meilleurs paramètres à l'aide de grid Search.
- L'utilisation de l'ACP pour réduire les dimensions et choisir seulement les caractéristiques dont les charges sont élevées, comme le montre la figure suivante 4.

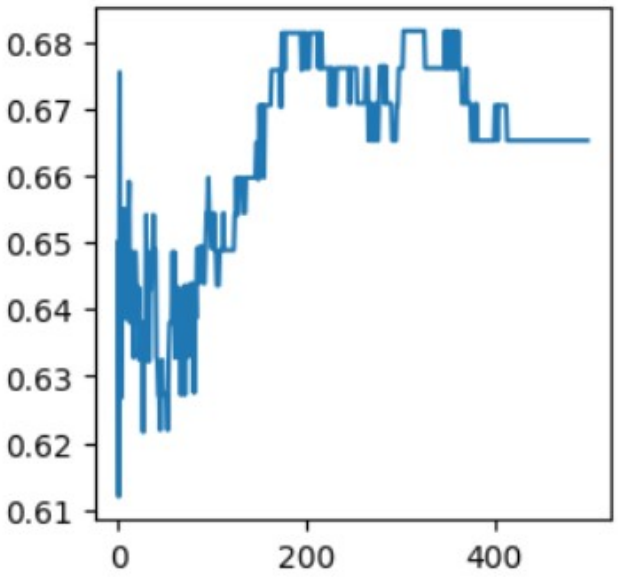

FIGURE 3 – Courbe d'accuracy en fonction du paramètre n_estimators

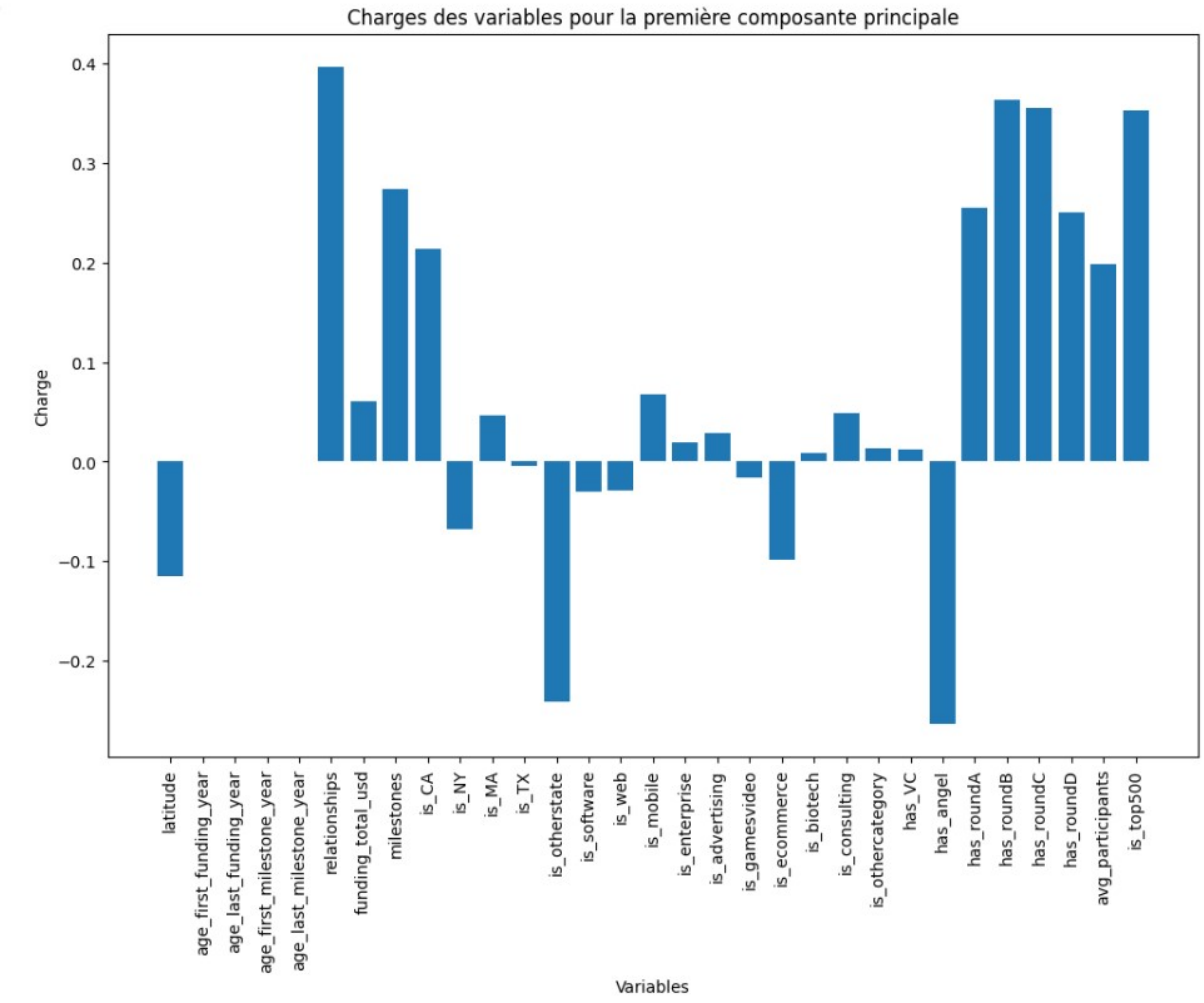

FIGURE 4 – Charges des variables pour la première composante principale

- Utiliser le RFE pour feature selection aussi en plus de l'ACP.
- Procéder avec des itérations lors de l'entraînement du modèle

5.8 Déploiement du modèle

Le déploiement d'un modèle d'apprentissage machine s'avère nécessaire pour predire le succes d'une startup ou d'une entreprise et offrir de nombreux avantages aux decideurs ce qui peut aider des entreprises, bureaux de conseils et économistes à prendre des décisions éclairées sur les demandes de leurs clients et savoir ou investir.

Il est important de noter que le déploiement d'un modèle d'apprentissage machine nécessite généralement des compétences en ingénierie logicielle, en gestion des données et en infrastructure informatique. De plus, il est crucial de maintenir et de mettre à jour régulièrement le modèle une fois qu'il est déployé pour garantir sa précision et son bon fonctionnement dans des conditions réelles.

Alors, le choix de l'environnement de deploiement est un aspect fondamental du processus de mise en place d'un modele de machine learning. Le deploiement sur site peut etre plus adapte aux entreprises qui ont des ressources informatiques et des competences techniques suffisantes pour gerer le modele en interne, et même pour celles qui ont des besoins en matiere de confidentialite ou de securite des donnees.

Mais aussi, pour de nombreuses entreprises, le deploiement sur le cloud est souvent plus pratique et plus economique, il offre egalement une plus grande flexibilite et evolutivite pour les entreprises. Les services de cloud computing tels que Amazon Web Services, Microsoft Azure et Google Cloud Platform proposent des solutions de deploiement de modeles de machine learning preconfigurees, avec des outils de developpement et de gestion de modeles faciles a utiliser.

En raison des contraintes cités auparavant, il nous a été difficile de déployer notre modèle.

6 Conclusion

Dans ce présent projet, nous avons mis en œuvre les compétences acquises dans le domaine de la data science et de l'intelligence artificielle afin de construire un modèle d'apprentissage machine capable de prédire le succès d'une startup.

Malgré les obstacles confrontés vu le manque d'expérience dans ce domaine là, loin de nous décourager, nous avons consacré beaucoup de notre temps à réunir les informations et les connaissances nécessaires pour atteindre l'objectif fixé pour ce projet.