BIEDER Bianca
BILLY Salomé

PYTHON FOR DATA ANALYSIS - PROJET 2021

Polish companies bankruptcy data

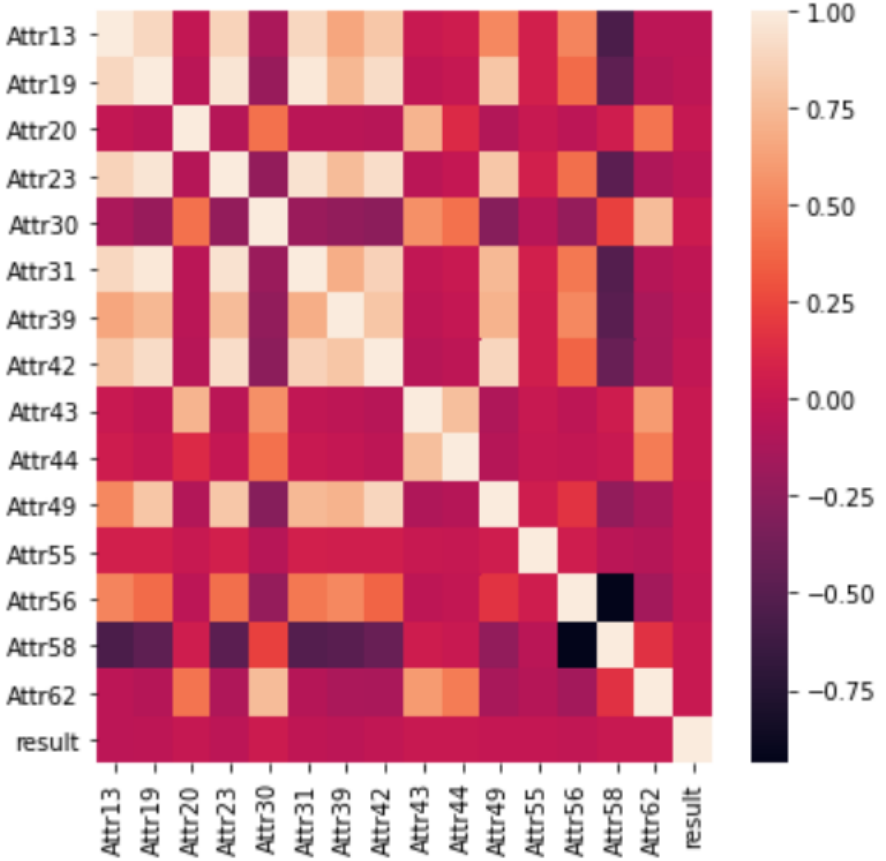
LA BASE DE DONNÉES

- La base de données mise à notre disposition sur le site UCI représente les faillites des entreprises Polonaises. L'étude a été réalisée entre 2000 et 2012 pour les entreprises qui ont fait faillite et entre 2007 et 2013 pour les entreprises qui sont encore en activité.
- Dans cette base de donnée, nous disposons de 5 différents datasets. Ces derniers représentent l'année à laquelle a été réalisée les prédictions de faillites 5 ans après le début des prédictions (Pour la première/troisième année on retrouve les chiffres associés aux différentes entreprises pendant cette année et la prédiction de faillite 5/3 ans après).
- Les différentes variables, au nombre de 64, présentes dans chacun des datasets représentent en globalité les chiffres/statistiques économiques des entreprises, la dernière variable (chiffre binaire) représente la faillite (1) de l'entreprise ou non (0).

LE PROBLÈME

- En étudiant les données et les variables disponibles pour cette étude, nous avons tout de suite dégagé une question évidente pour ce sujet, quelles sont les variables les chiffres moyens par attributs qui pourrait indiquer une future faillite ?
- Notre problème serait donc de pouvoir, par la suite, estimer si une entreprise a des risques de faire faillite en fonction de ses chiffres économiques et de l'année de l'étude.
- Tout d'abord nous avons décidé d'évaluer les corrélations entre chacune des variables et d'en sélectionner une quinzaine pour effectuer l'étude du problème. Plusieurs variables sont assez proches ou représentent des chiffres assez identiques, il a donc été possible d'en éliminer quelques-uns.

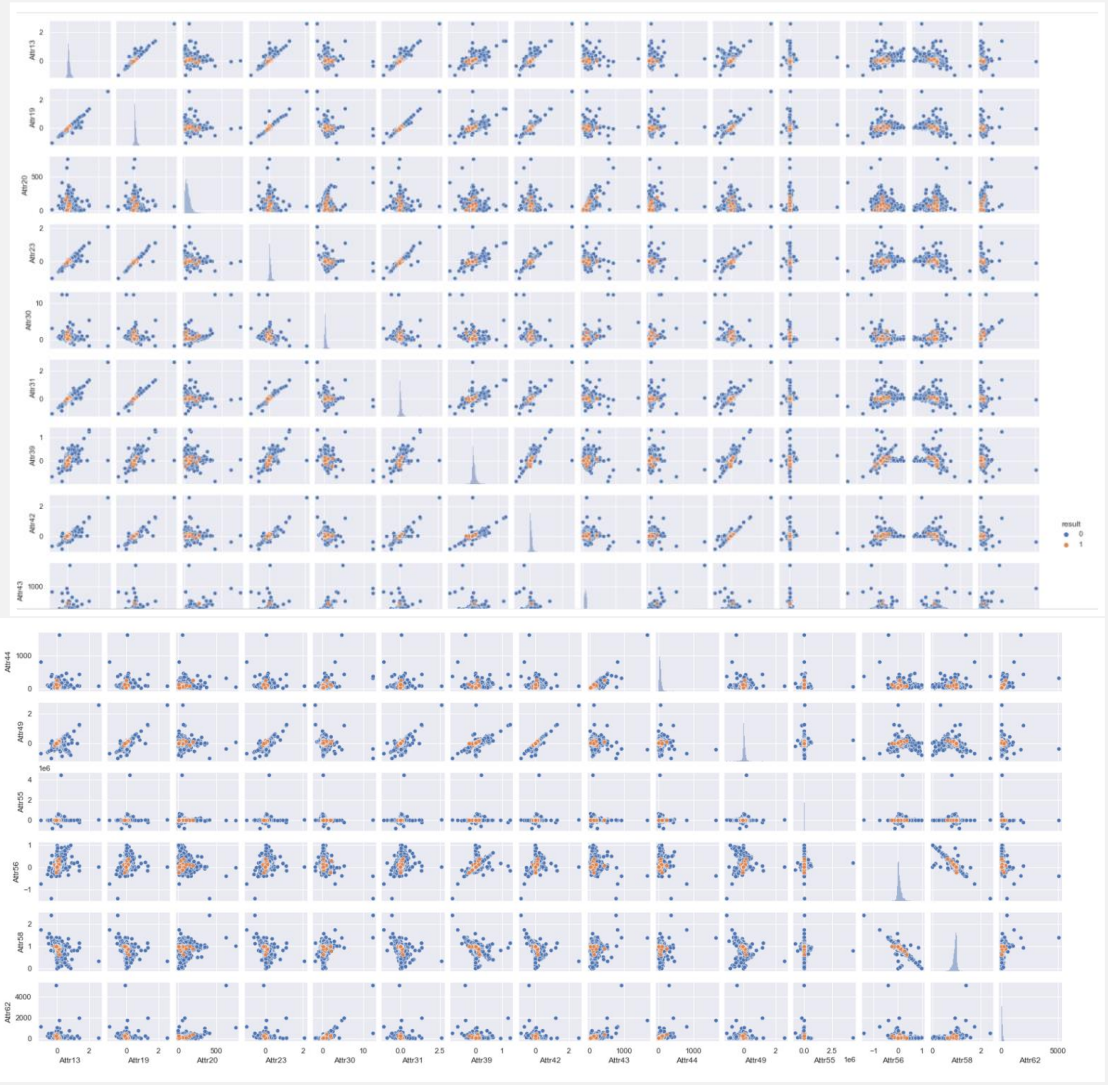
CHOIX DES VARIABLES ET CORRÉLATIONS ENTRE ELLES



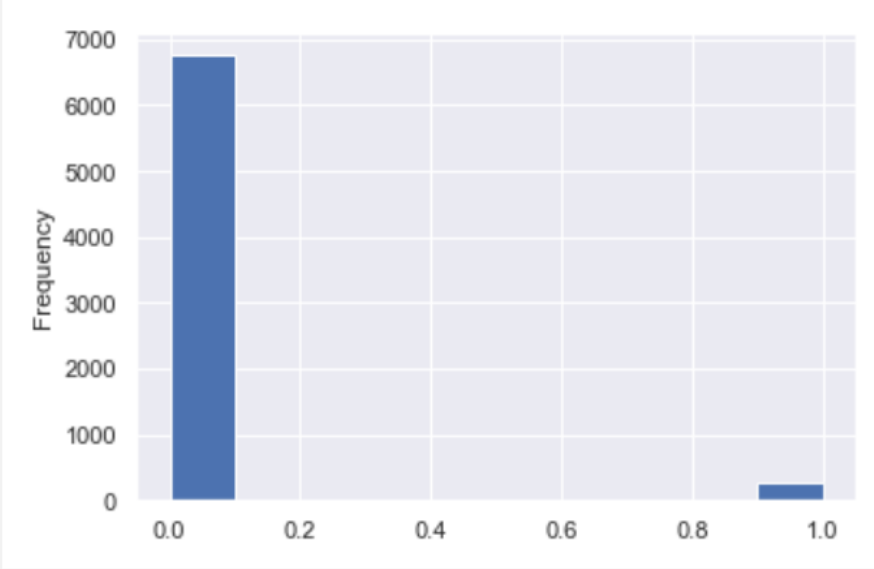
- Nous avons tout d'abord décidé d'utiliser un corrplot() de la librairie seaborn qui nous a permit de visualiser quelles variables étaient corrélées entre elles. Nous avons décidé de supprimer l'attribut 55 car la corrélation était nulle.

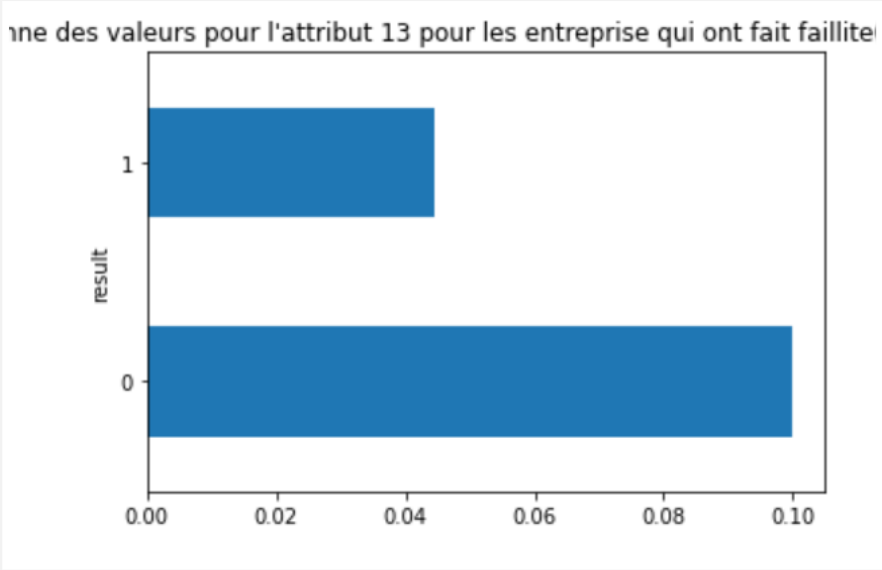
- Nous avons ensuite fait le choix de garder seulement quelques-unes de ces variables pour que l'étude soit plus poussée sur certaines variables précises. Nous avons donc conservé 14 de ces variables (Attr 13,19,20,23,30,31,39,42,43,44,49,56,58,62)

DATA-VISUALISATION: ANNÉE 1



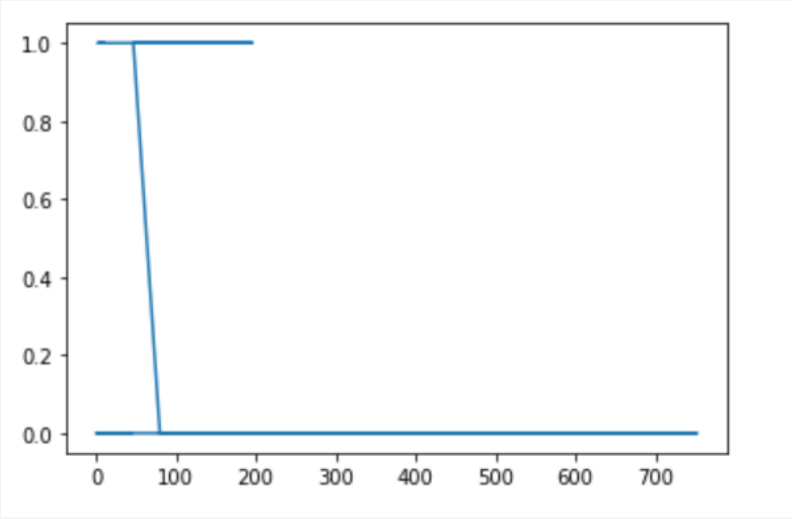
- Nous avons tout d'abord utilisé la librairie Matplotlib et Seaborn pour afficher des graphiques/tableaux représentant les variables en fonction du résultat (si l'entreprise fait faillite ou non). Nous avons affiché les graphiques pour chacune des variables conservées pour l'étude.
- Nous avons tout d'abord étudié les chiffres de la première année (year1), en utilisant les deux librairies citées ci-dessus notre but était de montrer l'influence de chacun des attribut sur le résultat.
- Le premier graphique que nous avons utilisé est le pairplot (seaborn), il nous a permis de visualiser les relations entre les attributs (nuages de points) mais aussi leur influence sur le résultat (histogramme sur la diagonale).

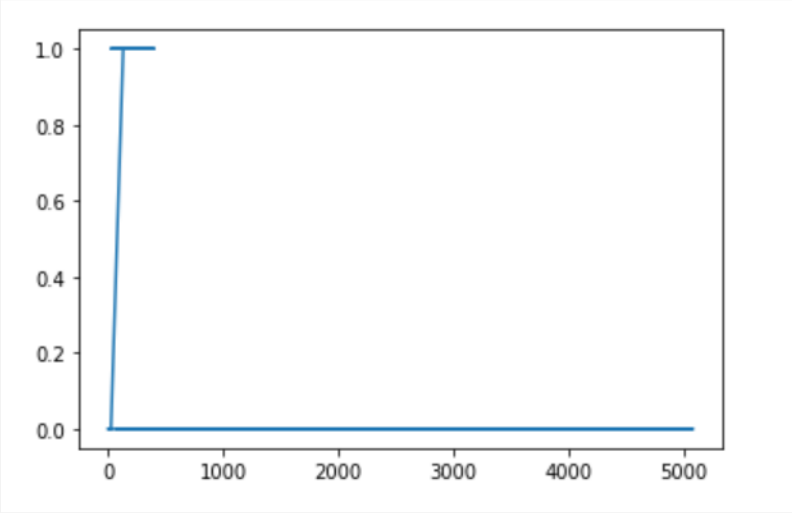




DATA-VISUALISATION: ANNÉE 1

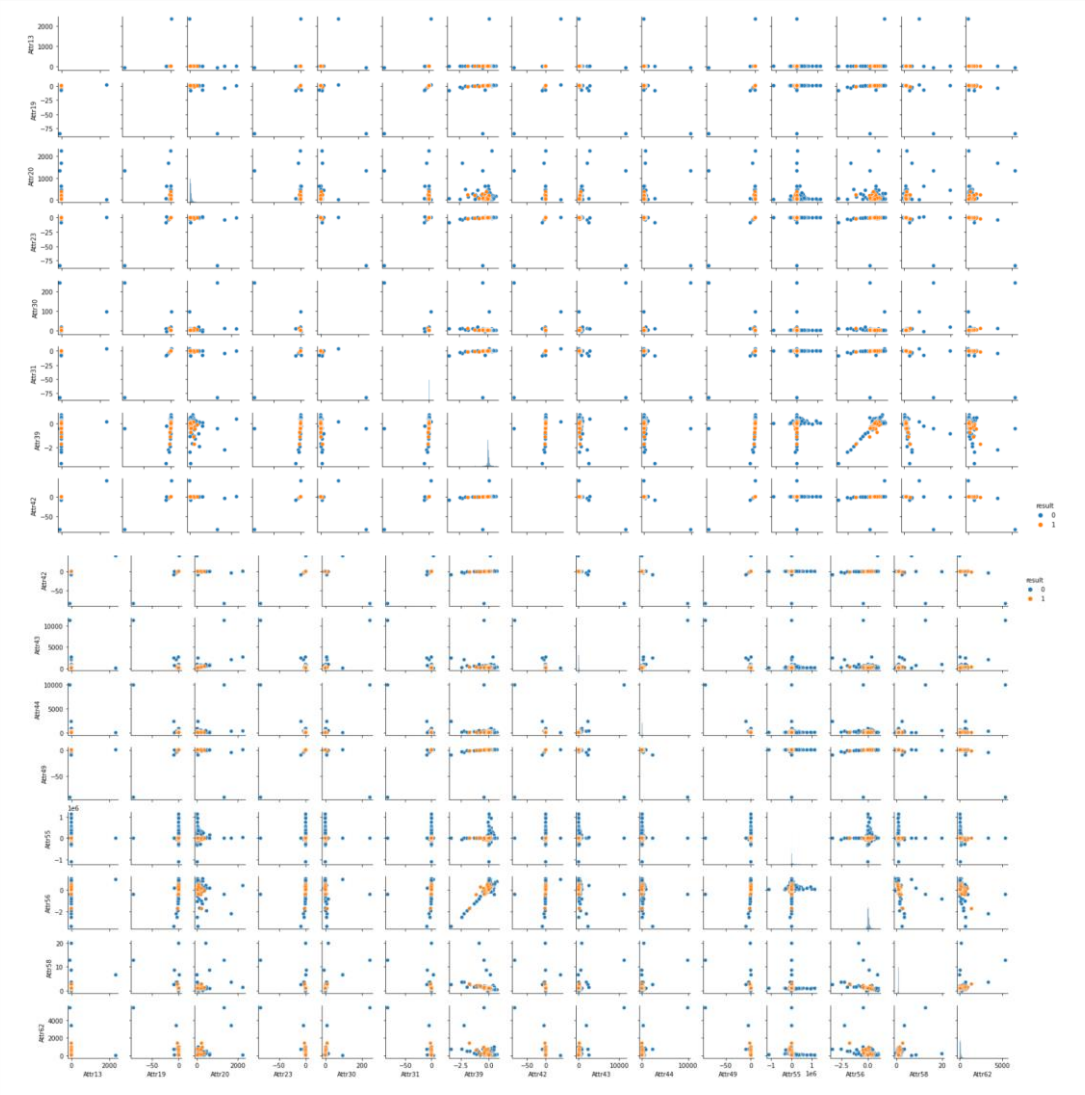
- Nous avons ensuite tracé un histogramme représentant le nombre de faillite ou non faillite cette première année. On remarque tout de suite que les entreprises ayant fait faillite reste dans un nombre très faible pour le moment.
- Afin de visualiser plus précisément l'influence de chaque variable sur le résultat nous avons affiché les tableaux des moyennes/écart-types des attributs en fonction de result. La différence de valeurs est observable pour tous les attributs, cependant, pour certains attributs la différence entre les deux résultats est très large ce qui montre leur influence comme (inventory * 365) / sales, (total liabilities - cash) / sales, rotation receivables + inventory turnover in days, (receivables * 365) / sales, (sales - cost of products sold) / sales.
- Plus précisément, nous avons représenté la moyenne des valeurs des attributs en fonction du résultat, on remarque que pour l'attribut « (gross profit + depreciation) / sales » une faillite est observé en moyenne pour des valeurs égales à 0.04.





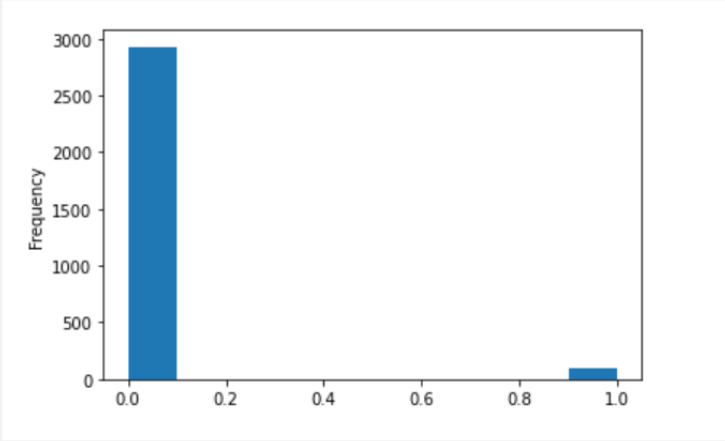
DATA-VISUALISATION: ANNÉE 1

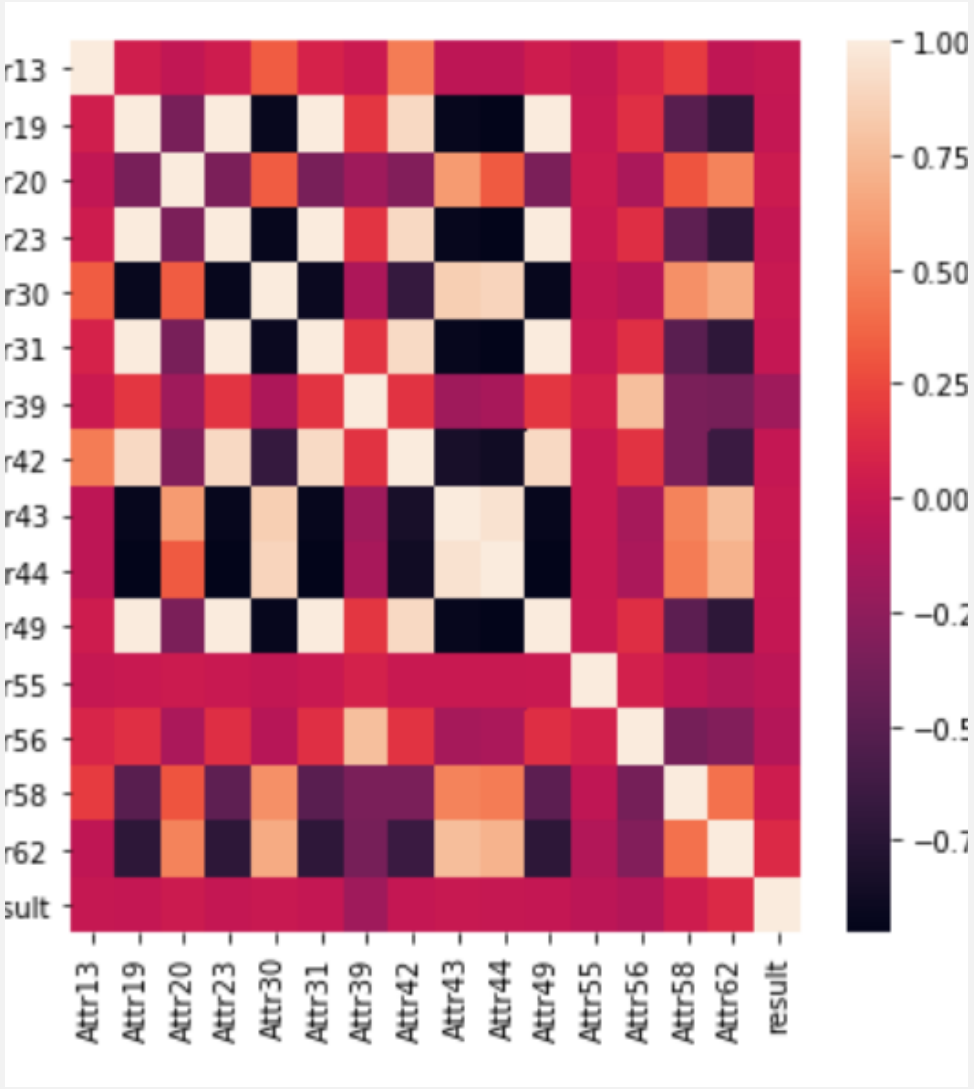
- Enfin nous avons décidé de tracer des graphiques qui représentent les valeurs des attributs en fonction du résultat. On remarque par exemple que pour des valeurs entre 0 et 200 pour l'attribut « (inventory * 365) / sales » l'entreprise peut avoir un risque de faire faillite.
- Pour l'attribut « (short-term liabilities *365) / sales » on remarque qu'une entreprise a plus de chance de faire faillite pour des faibles valeurs.
- Enfin nous avons observé de plus près les plot du pairplot représentant les lien entre les attributs, cela nous a permit de voir quels attributs ont une grandes influences sur la faillite d'une entreprise (si les points représentant une faillite sont bien isolés du reste des points). Par exemple, selon nos graphiques, on a remarqué que les attributs « gross profit + depreciation) / sales » et « gross profit / sales » permettent bien de délimiter les faillites des non-faillites.



DATA-VISUALISATION: ANNÉE 5

- Etant donné que nous disposions de 5 datasets différents représentant 5 année consécutives, nous avons décidé de réaliser une deuxième étude comme la précédente sur la 5ème afin de pouvoir mesurer la différence entre le début de l'étude et la fin. Le nombre de faillites restent toujours beaucoup plus faible que les entreprises qui poursuivent leur activité (comme on peut le voir sur l'histogramme).



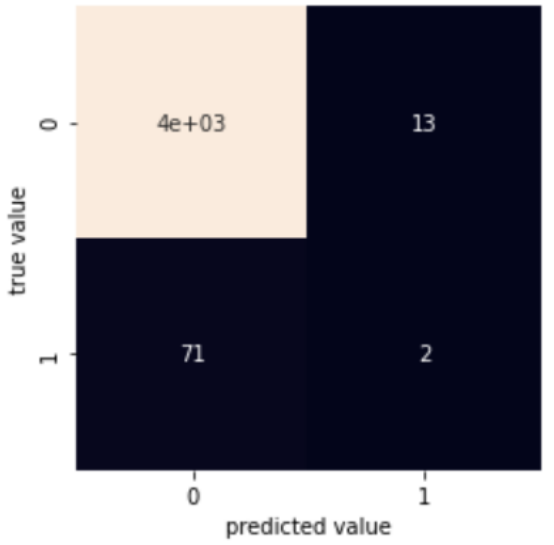


DATA-VISUALISATION: ANNÉE 5

- Le corrplot de l'année 5 est différent de la première année mais il montre que ce sont les mêmes attributs qui peuvent être conservés pour la suite de l'étude.
- Les graphiques tracés ont montré pour l'attribut « profit on sales / sales » pour des valeurs comprises entre -1.5 et 0, l'entreprise a un risque de faire faillite.
- De plus, pour des faibles valeurs de l'attribut « (inventory * 365) / sales » on remarque que les entreprises ont fait faillite.

MODÉLISATION

- Pour la partie de modélisation il nous était demandé d'utiliser la librairie scikit-learn, nous avons donc réaliser plusieurs modèles avec des méthodes de machine learning différents et des hyperparamètres modifiés: Linear Regression, Gaussian NB, KNN et Random Forest.
- Tout d'abord nous avons utilisé la méthode Linear Regression avec laquelle nous avons créé un modèle, pour tester cette méthode nous avons donc décidé de prédire le résultat pour la deuxième année (en chargeant les données de year2 et en supprimant les mêmes colonnes que pour la première année).
- La méthode de Gaussian NB n'est pas le bon modèle a utiliser, en effet le score de précision est très faible, même en changeant les hyperparamètres (0.0421172453044963, 0.05122367672168469).
- La méthode KNN nous a retourné des scores de prédictions très bon, de mieux en mieux lorsque l'on réduit le nombre de voisins.
- Enfin la méthode que nous avons conservé est celle de Random Forest qui a donné un bon score et qui a prédit les valeurs de la deuxième année avec peu de fautes. Nous en avons conclu que le très faible nombre de faillite la première année est ce qui complique la prédiction pour les années suivantes, par faute d'apprentissage.



MODÉLISATION – GRILLE DE RECHERCHE

Méthode utilisée	Hyperparamètres	Accuracy Score
Linear Regression	fit_intercept=True	
Gaussian NB	random_state=120	0.042117245304496 3
Gaussian NB	random_state=1	0.051223676721684 69
KNN	n_neighbors=4	0.962288316493525
KNN	n_neighbors=1	0.999573075281058 8
Random Forest	random_state=0	0.979452054794520 6
Random Forest	random_state=15	0.977739726027397 2

TRANSFORMATION EN API

- Une fois le modèle Random Forest validé nous avons créé une API sur Flask afin de pouvoir visualiser notre étude sur le serveur. Le but de cet API est de pouvoir entrer une valeurs pour chaque attribut demandé et l'API pourra analyser ces valeurs et prédire si l'entreprise risque de faire faillite ou non. Pour ceci nous avons créé deux fichier python (app et model) et un fichier pickle vide.
- Pour cela, dans le fichier app.py on a choisit le dataset 'year2' dont nous avons gardé les 15 attributs suivants : (gross profit + depreciation) / sales, gross profit / sales, (inventory * 365) / sales, net profit / sales, (total liabilities - cash) / sales, (gross profit + interest) / sales, profit on sales / sales, profit on operating activities / sales, rotation receivables + inventory turnover in days, receivables * 365) / sales, EBITDA (profit on operating activities - depreciation) / sales, working capital, (sales - cost of products sold) / sales, total costs /total sales, (short-term liabilities *365) / sales

TRANSFORMATION EN API

- A partir de ces données, nous utilisons la régression linéaire et nous entrainons notre modèle de prédiction. Enfin nous sauvegardons ce modèle dans un fichier model.pickle. Le fichier app.py permet d'initialer API avec flask. Il permettra de lire les cases remplies par utilisateur et retournera le résultat (risque de faillite ou non en s'appuyant sur le modèle créé précédemment).
- Enfin le fichier index.html permet d'afficher les différentes cases à remplir par l'utilisateur, les boutons et les titres sur la page d'accueil.

CONCLUSION

- Les différents graphiques utilisés on peut montrer les valeurs des plusieurs chiffres économiques des entreprises à partir desquelles elles peuvent risquer la faillite. Aux vues des résultats du pairplot on a pu voir que la prédiction était assez compliquée, les points représentant les faillites sont très peu isolés du reste des points, ce qui montre que les différents attributs ne permettent pas tout seul de déterminer une faillite. Chacun des chiffres peuvent donc indiquer le seuil ou les entreprises risques d'avoir des problèmes économiques.
- Avec la librairie sickit learn nous avons pu voir que la méthode de Random Forest était la meilleure mais les scores montrent tout de même quelques failles (des erreurs de fausse prédiction). Cet prédiction est compliqué car sur le nombre total d'entreprises, très peu sont en faillites, la machine n'a donc pas beaucoup de chiffres d'apprentissage pour cela: ce qui amène a des fausses prédictions.