ENSEA Projet S9

# Comparaison d'algorithmes de découverte causale sur différents jeux de données

Anis AFLOU et Martin GERVAIS

## Remerciements

Nous tenons à remercier M. Vassilis et M. Vareille pour leur encadrement, leurs conseils et leurs remarques constructives tout au long de ce projet. Leurs retours nous ont permis d'affiner notre méthodologie, de mieux structurer notre démarche expérimentale et d'adopter une approche plus rigoureuse vis-à-vis des algorithmes de découverte causale.

## Résumé

La découverte causale vise à inférer des relations de cause à effet à partir de données observationnelles, généralement représentées sous forme de graphes causaux. En pratique, l'évaluation des algorithmes de découverte causale reste difficile, car ces méthodes reposent sur des hypothèses fortes et sont souvent appliquées à des jeux de données et des protocoles expérimentaux différents.

Dans ce contexte, ce projet propose une étude comparative de plusieurs algorithmes issus de familles méthodologiques distinctes (*Peter-Clark, Hill-Climbing et DirectLiNGAM*), appliqués à trois jeux de données réels de natures différentes. Un protocole expérimental commun est mis en place, en adaptant le prétraitement aux hypothèses de chaque algorithme.

Les graphes causaux obtenus sont évalués à l'aide de la puissance prédictive des Markov Boundaries associées à la variable cible.

# Sommaire

# I. Introduction

Les méthodes de machine learning sont aujourd'hui utilisées pour analyser des données complexes et réaliser des prédictions dans différents domaines d'études. Ces approches reposent toutefois principalement sur l'identification de corrélations statistiques et ne permettent pas, en général, d'établir des relations de cause à effet entre les variables observées.

La découverte causale vise à dépasser cette limitation en cherchant à inférer la structure causale sous-jacente aux données observationnelles. Cependant, son application en pratique soulève plusieurs difficultés. Les jeux de données réels sont souvent hétérogènes, combinant des variables continues, discrètes, catégorielles ou temporelles, tandis que les algorithmes de découverte causale reposent sur des hypothèses spécifiques concernant la nature des données et les relations qu'elles entretiennent. En conséquence, l'application de différents algorithmes sur un même jeu de données peut conduire à des graphes causaux sensiblement différents.

Par ailleurs, bien que de nombreux algorithmes de découverte causale aient été proposés dans la littérature, leur comportement pratique reste difficile à évaluer. Les méthodes sont fréquemment testées sur des jeux de données, des métriques et des protocoles expérimentaux distincts, ce qui limite les comparaisons directes et systématiques. Il existe ainsi un manque d'évaluations comparatives menées dans un cadre expérimental commun.

L'objectif de ce projet est d'étudier et de comparer plusieurs algorithmes de découverte causale issus de familles méthodologiques distinctes, en les évaluant dans des conditions expérimentales homogènes et sur des jeux de données de natures différentes. Un protocole expérimental commun est mis en place, dans lequel chaque algorithme est appliqué en respectant ses hypothèses théoriques. Les graphes causaux obtenus sont ensuite analysés d'un point de vue structurel et évalués à l'aide de la puissance prédictive des Markov Boundaries associées à la variable cible.

# II. Notions de causalité et découverte causale

## Causalité et relations de cause à effet

La causalité désigne une relation dans laquelle une variable (la cause), produit un effet mesurable sur une autre variable (l'effet). Contrairement à une corrélation statistique, la relation causale implique ce que l'on appelle une asymétrique directionnelle, c'est-à-dire que la cause précède l'effet et en est l'origine.

Ainsi, observer une dépendance statistique entre deux variables ne suffit pas à conclure l'existence d'un lien causal. En effet, deux variables peuvent être corrélées sans qu'aucune n'influence directement l'autre. La causalité cherche justement à distinguer ces situations en identifiant ce qui relient ces variables entre elles.

## Représentation graphique de la causalité

Les relations causales sont représentées à l'aide de graphes causes dirigés acycliques (autrement appelé DAG).

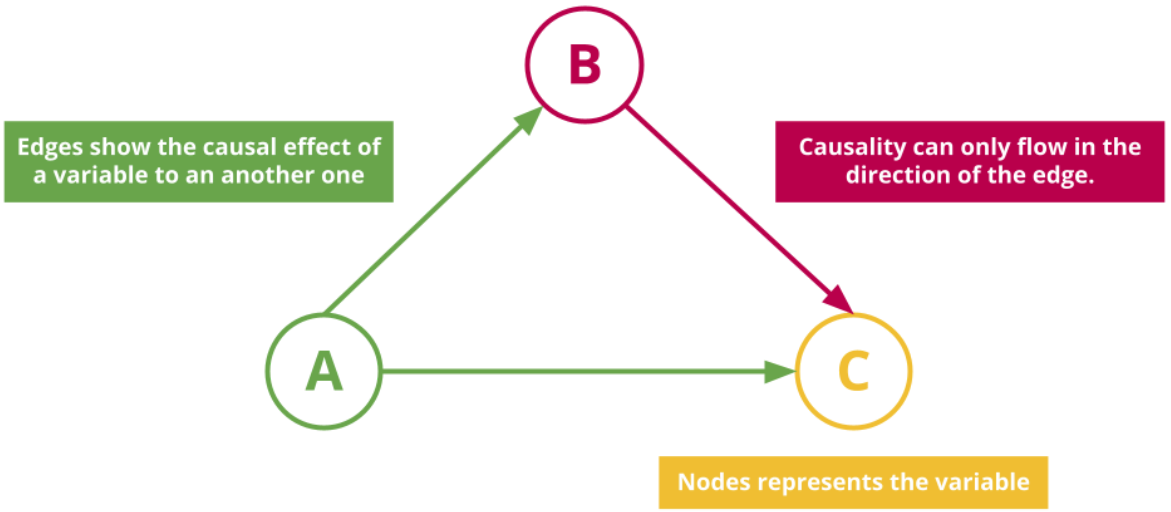



*Figure 1 : Exemple explicatif du DAG*

Dans ces graphes, les nœuds représentent les variables et les arêtes orientées les relations de causes à effet.
Une arête orientée de la variable **A** à la variable **B** signifie que **A** à une influence causale sur **B**. La causalité ne peut circuler que dans le sens de la flèche. Si l'on observe une absence d'arête entre deux variables cela implique l'absence de causalité de ces deux dernières.

# III. Présentation des datasets

## Présentation du jeu de données Lending Club

### 1) Description générale du jeu de données

Le jeu de données Lending Club provient d'une plateforme américaine de prêt entre particuliers. Cette plateforme met en relation des investisseurs souhaitant prêter de l'argent et des emprunteurs cherchant à financer différents projets personnels. Les investisseurs perçoivent un rendement sous forme d'intérêts tandis que les emprunteurs bénéficient de taux plus avantageux que ceux proposés par les institutions financières traditionnelles.

La dataset ici utilisé regroupe l'ensemble de prêts accordés par Lending Club sur la période 2007-2015. Il s'agit d'un jeu de données réel, trouvé sur kaggle.com.

### 2) Taille et structure des données

Le fichier loan.csv contient 890 000 observations et 75 variables. Ces dernières peuvent décrire à la fois les caractéristiques du prêt, le profil de l'emprunteur et l'historique de crédit. Afin de simplifier l'étude, pour des raisons de temps de calcul et de lisibilité nous avons réduit de 75 variables à 11.

Ainsi, chaque observation est associée à un statut de remboursement du prêt (*loan status*), indiquant si le prêt est "*Fully Paid*", "*Current*", "*Late*" ou "*Charged Off*".

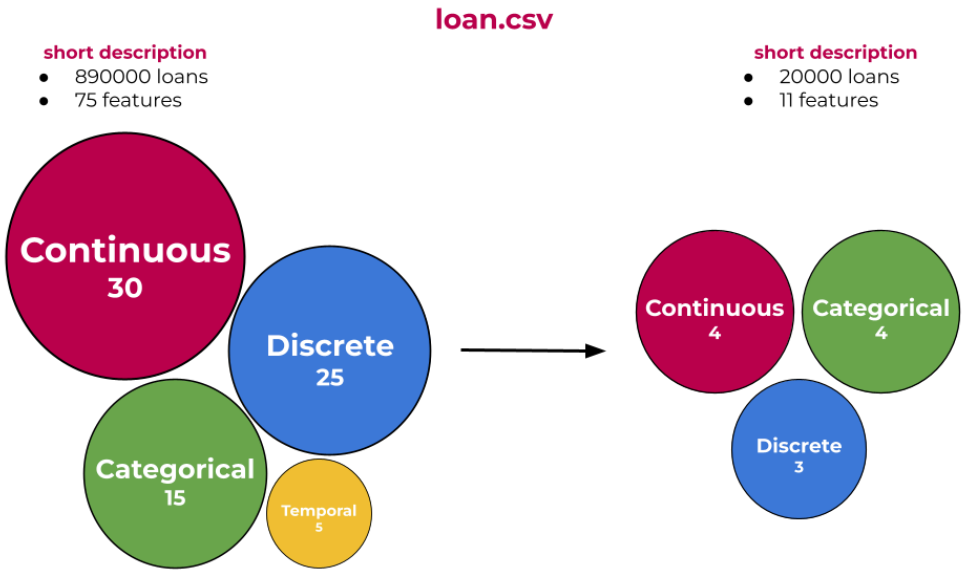



*Figure 2 : bubble charts de loan.csv*

### 3) Loan.csv – Description des variables sélectionnées

| loan.csv, description des variables sélectionnées | | |
|---|---|---|
| **nom** | **courte description** | **type de données** |
| **loan_status** | **Statut final du prêt indiquant s'il est remboursé, en cours ou en défaut. Variable cible de l'étude.** | **catégorielle (cible)** |
| annual_inc | Revenu annuel déclaré par l'emprunteur au moment de la demande de prêt. | numérique continu |
| dti | Ratio d'endettement correspondant au rapport entre les charges mensuelles de dettes et le revenu mensuel. | numérique continu |
| loan_amnt | Montant total du prêt demandé par l'emprunteur. | numérique continu |
| revol_util | Taux d'utilisation du crédit renouvelable par rapport au crédit disponible. | numérique continu |
| delinq_2yrs | Nombre d'incidents de paiement survenus au cours des deux années précédant la demande de prêt. | numérique entier |
| open_acc | Nombre de comptes de crédit ouverts au nom de l'emprunteur. | numérique entier |
| total_acc | Nombre total de comptes de crédit détenus par l'emprunteur, ouverts ou fermés. | numérique entier |
| term | Durée du prêt exprimée en mois (par exemple 36 ou 60 mois). | catégorielle |
| home_ownership | Statut de logement de l'emprunteur (propriétaire, locataire, hypothèque, autre). | catégorielle |
| emp_length | Ancienneté professionnelle de l'emprunteur, exprimée en tranches d'années. | catégorielle |

*Figure 3 : description des variables sélectionnées pour loan.csv*

## Présentation du jeu de données GiveMeSomeCredit

### 1) Description générale du jeu de données

Le jeu de données GiveMeSomeCredit.csv provient d'une compétition Kaggle organisée par Credit Fusion dont l'objectif est d'améliorer les méthodes de credit scoring. Le problème consiste à prédire la probabilité qu'un individu rencontre une situation de détresse financière grave au cours des deux années suivant l'observation.

La variable cible, *SeriousDlqin2yrs* est binaire et un indique si un emprunteur a connu un défaut de paiement sérieux (retard supérieur à 90 jours) dans les deux années.

### 2) Taille et structure des données

Le jeu de données contient 150 000 observations, chacune correspond à un individu et 11 variables explicatives. La structure des données est la suivante :

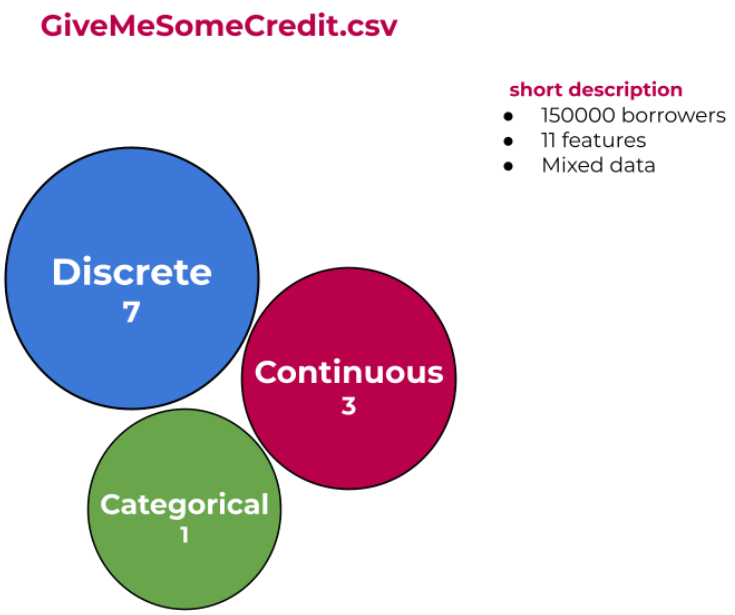



*Figure 4 : bubble charts pour GiveMeSomeCredit.csv*

### 3) Description des variables

| GiveMeSomeCredit.csv, description des variables | | |
|---|---|---|
| **nom** | **courte description** | **type de données** |
| **SeriousDlqin2yrs** | **Indique si l'individu a connu un retard de paiement sérieux (≥ 90 jours) dans les deux années suivant l'observation. Variable cible.** | **binaire (cible)** |
| RevolvingUtilizationOfUnsecuredLines | Ratio d'utilisation des lignes de crédit non garanties. | numérique continu |
| Age | Âge de l'emprunteur en années. | numérique entier |
| NumberOfTime30-59DaysPastDueNotWorse | Nombre de retards de paiement compris entre 30 et 59 jours. | numérique entier |
| DebtRatio | Ratio d'endettement correspondant au total des charges mensuelles divisé par le revenu mensuel. | numérique continu |
| MonthlyIncome | Revenu mensuel déclaré par l'emprunteur. | numérique continu |
| NumberOfOpenCreditLinesAndLoans | Nombre total de lignes de crédit ouvertes et de prêts en cours. | numérique entier |
| NumberOfTimes90DaysLate | Nombre de retards de paiement de 90 jours ou plus. | numérique entier |
| NumberRealEstateLoansOrLines | Nombre de prêts immobiliers et de lignes de crédit associées à l'immobilier. | numérique entier |
| NumberOfTime60-89DaysPastDueNotWorse | Nombre de retards de paiement compris entre 60 et 89 jours. | numérique entier |
| NumberOfDependents | Nombre de personnes à charge dans le foyer de l'emprunteur. | numérique entier |

*Figure 5 : description des variables pour GiveMeSomeCredit.csv*

## Présentation du jeu de données OnlineShoppersIntentions

### 1) Description générale du jeu de données

Le jeu de données OnlineShoppersIntention.csv décrit le comportement de navigation d'utilisateurs sur un site de commerce en ligne. L'objectif est de prédire si une session se conclut par un achat, représenté par la variable cible binaire Revenue.

### 2) Taille et structure des données

Le jeu de données contient 12 330 observations, chacune correspondant à une session utilisateur, et 18 variables explicatives avec la variable cible Revenue.

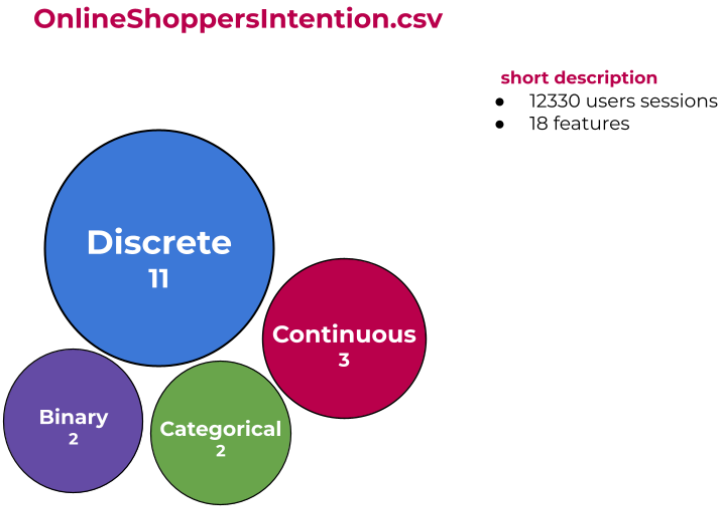



*Figure 6 : bubble chart pour OnlineShoppersIntention.csv*

### 3) Description des variables

| OnlineShoppersIntention.csv, description des variables | | |
|---|---|---|
| **nom** | **courte description** | **type de données** |
| Administrative | Nombre de pages administratives visitées durant la session. | numérique entier |
| Administrative_Duration | Temps total passé sur les pages administratives. | numérique entier |
| Informational | Nombre de pages informatives consultées. | numérique entier |
| Informational_Duration | Temps total passé sur les pages informatives. | numérique entier |
| ProductRelated | Nombre de pages liées aux produits visitées. | numérique entier |
| ProductRelated_Duration | Temps total passé sur les pages produits. | numérique continu |
| BounceRates | Taux de rebond moyen des pages visitées. | numérique continu |
| ExitRates | Probabilité moyenne qu'une page soit la dernière visitée lors d'une session. | numérique continu |
| PageValues | Valeur moyenne des pages visitées avant une transaction. | catégorielle |
| SpecialDay | Indicateur de proximité d'un événement commercial particulier. | numérique entier |
| Month | Mois au cours duquel la session a eu lieu. | numérique entier |
| OperatingSystems | Système d'exploitation utilisé durant la session. | numérique entier |
| Browser | Navigateur utilisé durant la session. | numérique entier |
| Region | Région géographique de l'utilisateur. | numérique entier |
| TrafficType | Source du trafic menant à la session. | numérique entier |
| VisitorType | Type de visiteur (nouveau, récurrent, autre). | catégorielle |
| Weekend | Indique si la session a eu lieu durant un week-end. | binaire |
| **Revenue** | **Indique si la session s'est conclue par un achat. Variable cible.** | **binaire (cible)** |

*Figure 7 : description des variables pour OnlineShoppersIntention.csv*

# IV. Méthodologie & pipeline expérimentale

## Vue d'ensemble du pipeline expérimentale

La Figure 2 présente le pipeline expérimental mis en place dans ce projet. Il décrit les différentes étapes suivies depuis l'importation des jeux de données jusqu'à l'évaluation des graphes causaux inférés.

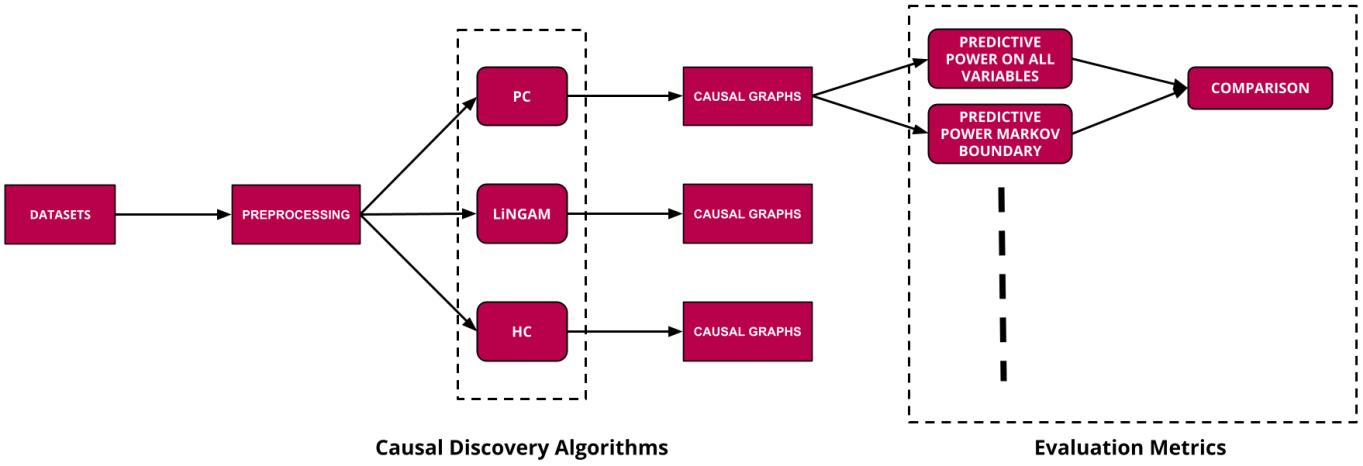



*Figure 8 : Pipeline expérimentale*

De manière générale, le pipeline se compose des étapes suivantes

1. Sélection et importation des données
2. Prétraitement en fonction des hypothèses de chaque algorithme
3. Découverte causale et inférence des graphes
4. Evaluation des graphes causaux à l'aide de métriques

Cette organisation permet d'appliquer chaque algorithme dans des conditions expérimentales homogènes, tout en tenant compte de leurs contraintes spécifiques lors des étapes de prétraitement.

## Sélection des variables

Afin de limiter la complexité du problème et de faciliter l'interprétation des graphes causaux obtenus, une sélection de variables est effectuée en amont de la découverte causale. Cette étape permet de réduire la dimension des jeux de données tout en conservant les variables les plus pertinentes pour l'analyse. Les variables sont retenues de manière intuitive et sont présentées dans la section précédente.

## Prétraitement et encodage des données

Le prétraitement de nos données est donc, comme nous l'avons compris, essentiel. En effet, il vise à rendre les jeux de données compatibles avec les algorithmes de découverte causale considérés, tout en préservant autant que possible l'information contenue dans les variables.

Les variables ordinales sont encodées à l'aide d'un encodage "naïf" respectant leur ordre naturel. L'utilisation d'un encodage one-hot est évitée car elle peut entraîner une augmentation importante de la dimension des données et perturber l'inférence des graphes causaux.

## Prétraitement spécifique aux algorithmes

Chaque algorithme de découverte causale est appliqué sur une version de jeu de données conformes aux hypothèses suivantes :

- **Algorithmes basés sur des contraintes (PC) :**
  - Appliqués sur des données mixtes
- **Algorithmes basés sur des scores (HC)**
  - Nécessitent des données discrètes
- **Modèles causaux structurels (LiNGAM)**
  - Uniquement sur des variables numériques continues

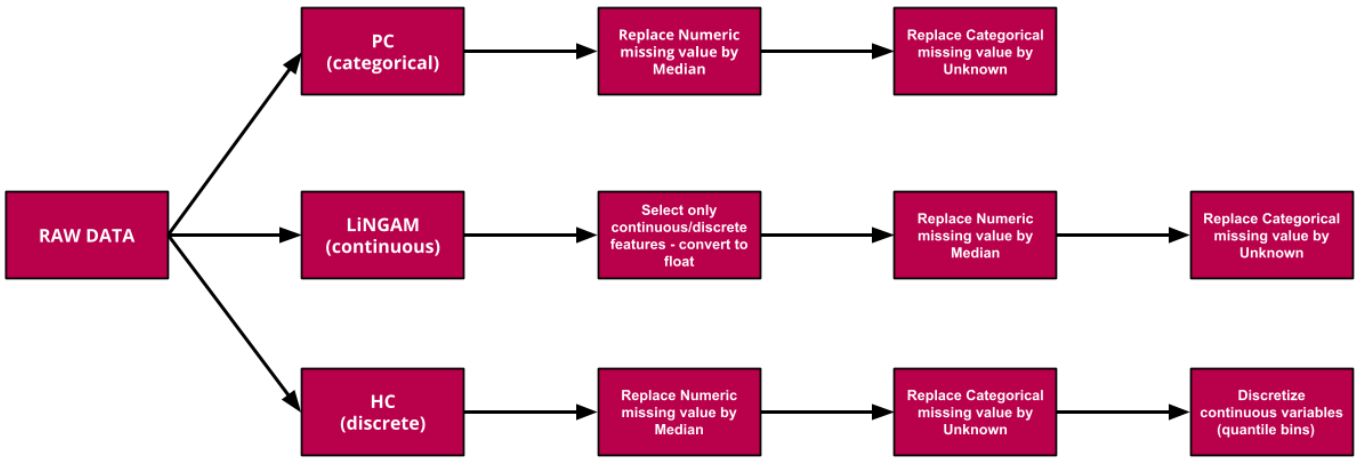



*Figure 9 : Data preprocessing*

## Découverte causale

Une fois les données préparées, chaque algorithme est appliqué afin d'inférer un graphe causal. Le graphe obtenu représente les relations de dépendance causale supposées entre les variables sélectionnées pour le jeu de données considérées.

Cette étape est réalisée indépendamment pour chaque algorithme et chaque jeu de données.

## Evaluation des graphes causaux

En l'absence de graphe causal de référence, l'évaluation des graphes inférés repose sur des métriques indirectes. Pour chaque graphe causal, la Markov Blanket (Parents, Enfants, Parents des enfants) de la variable cible est extraite. Un modèle prédictif est ensuite entraîné à partir des variables appartenant à cette Markov Boundary et sa performance est comparée à celle d'un modèle entraîné à partir de l'ensemble des variables.

La métrique du Structural Hamming Distance (SHD) n'est pas utilisée. En effet, elle nécessite la connaissance d'un graphe causal de référence qui est indisponible pour les jeux de données considérés.

# V. Description des algorithmes de découvertes causales

## Les familles d'algorithmes considérées

Dans ce projet nous comparons trois grandes familles d'algorithmes de découverte causale qui sont les suivantes :

- Les algorithmes basés sur la contrainte
- Les algorithmes basés sur des scores
- Les modèles causaux structurels

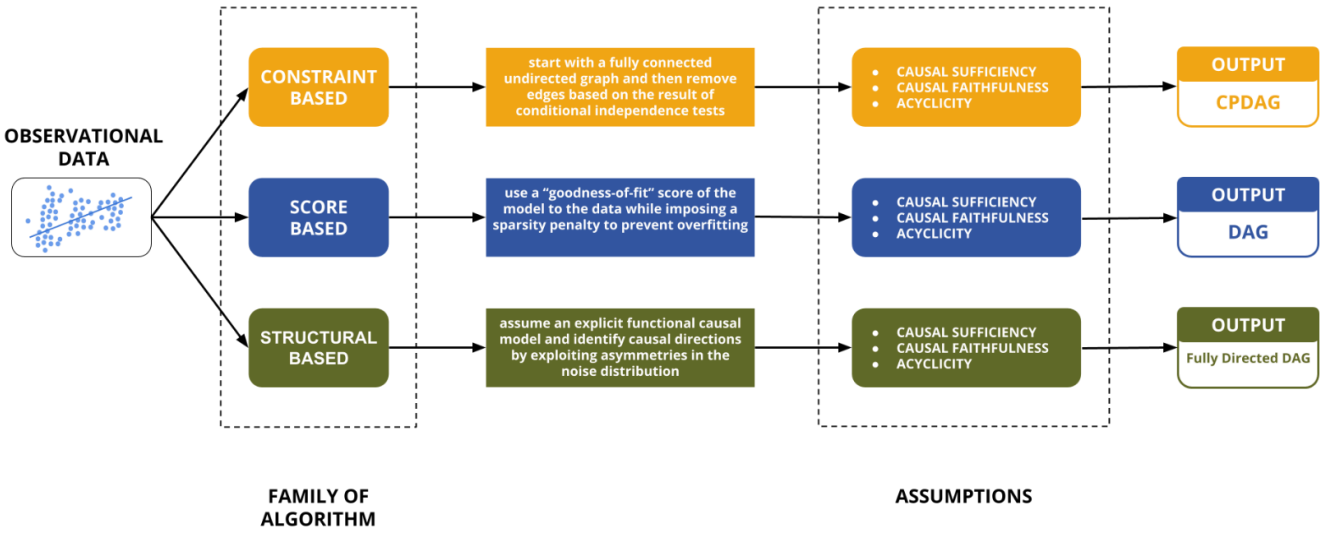



*Figure 10 : Les trois familles d'algorithmes choisies*

## Algorithme PC (méthodes basées sur la contrainte)

L'algorithme PC (*Peter-Clark*) appartient à la famille des méthodes basées sur des contraintes conditionnelles. Son principe de fonctionnement repose sur le fait que deux variables ne sont reliées par une arête dans un graphe causal que si elles ne peuvent être rendues indépendantes conditionnellement à un ensemble d'autres variables.

L'algorithme débute par un graphe complètement connecté, puis supprime progressivement les arêtes en effectuant des tests d'indépendance conditionnelle. Cela afin d'obtenir un CPDAG (Completed Partially Directed Acyclic Graph), représentant une de graphe causal.

Dans ce projet, nous utilisons un test d'indépendance basé sur la statistique de Pillai avec un niveau de significativité fixé à 5 % qui correspond à la sélectivité de l'algorithme. Plus le niveau de significativité est élevé, plus l'algorithme crée des liens. L'algorithme PC peut être appliqué à des jeux de données mixtes incluant à la fois des variables discrètes et continues.

**Pourquoi utiliser Pillai ?**

Le choix de l'hyper paramètre Pillai repose sur son fonctionnement. Celui-ci effectue ses tests d'indépendances causales en évaluant l'influence entre deux variables conditionnellement à une autre variable à l'aide de boosting. Boosting qui prend en charge des données mixtes ce qui nous permet d'éviter un encodage ou une discrétisation des données qui rendraient les résultats plus "bruitées".

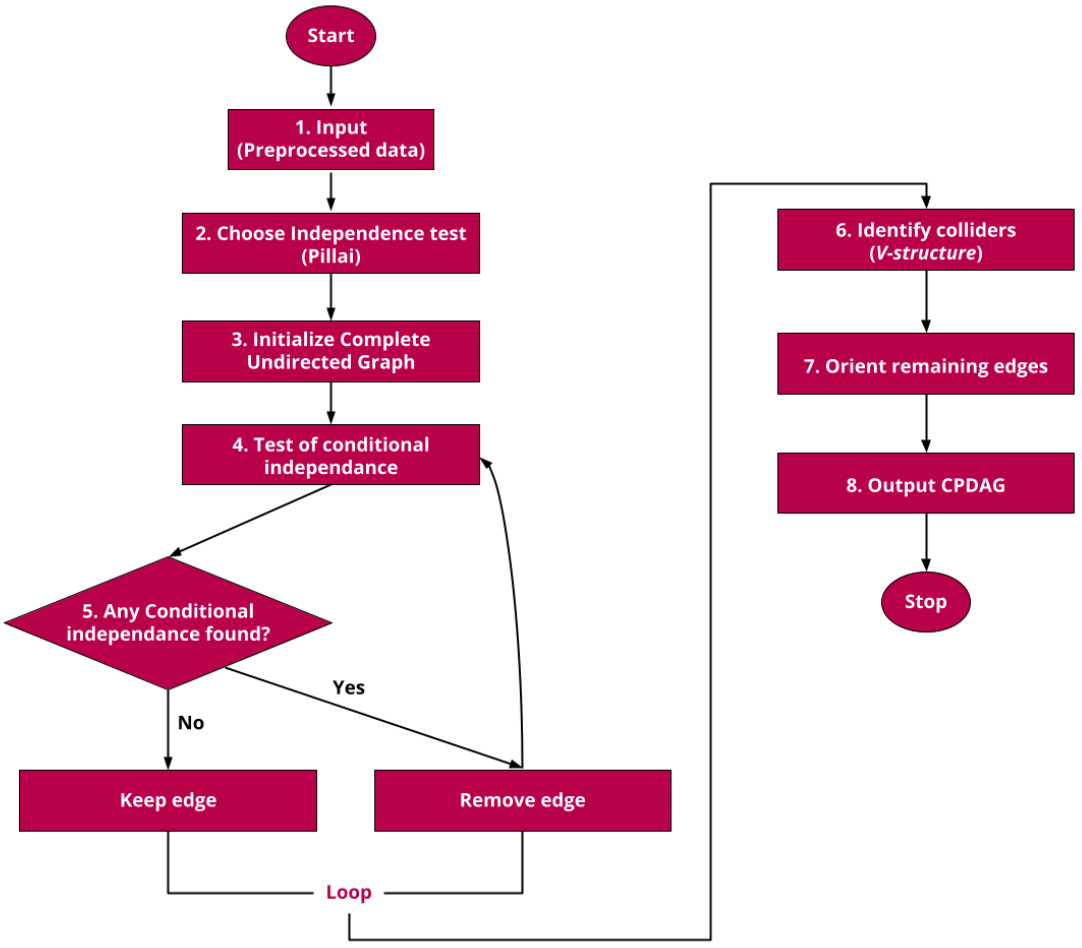



*Figure 11 : Flowchart du fonctionnement de PC*

## Algorithme HC (méthodes basées sur le score)

L'algorithme HC (*Hill-Climbing*) appartient à la famille des méthodes basées sur l'optimisation d'un score global. Il va donc chercher à identifier le graphe causal qui maximise un critère de qualité, dans notre projet il s'agit du Bayesian Information Criterion (BIC).

**Pourquoi utiliser BIC ?**
Car le BIC prend en compte des données catégorielles, de plus il est assez restrictif (plus restrictive que ces concurrents) donc intéressant pour ne pas avoir trop de liens

À partir d'un graphe initial l'algorithme explore un voisinage de graphes obtenus par ajout, suppression ou inversion d'arêtes et conserve la modification améliorant le score. Ce processus est répété jusqu'à convergence vers un optimum local.

Cependant, cet algorithme ne prend en compte uniquement des données discrètes, nécessitant donc une discrétisation des données.

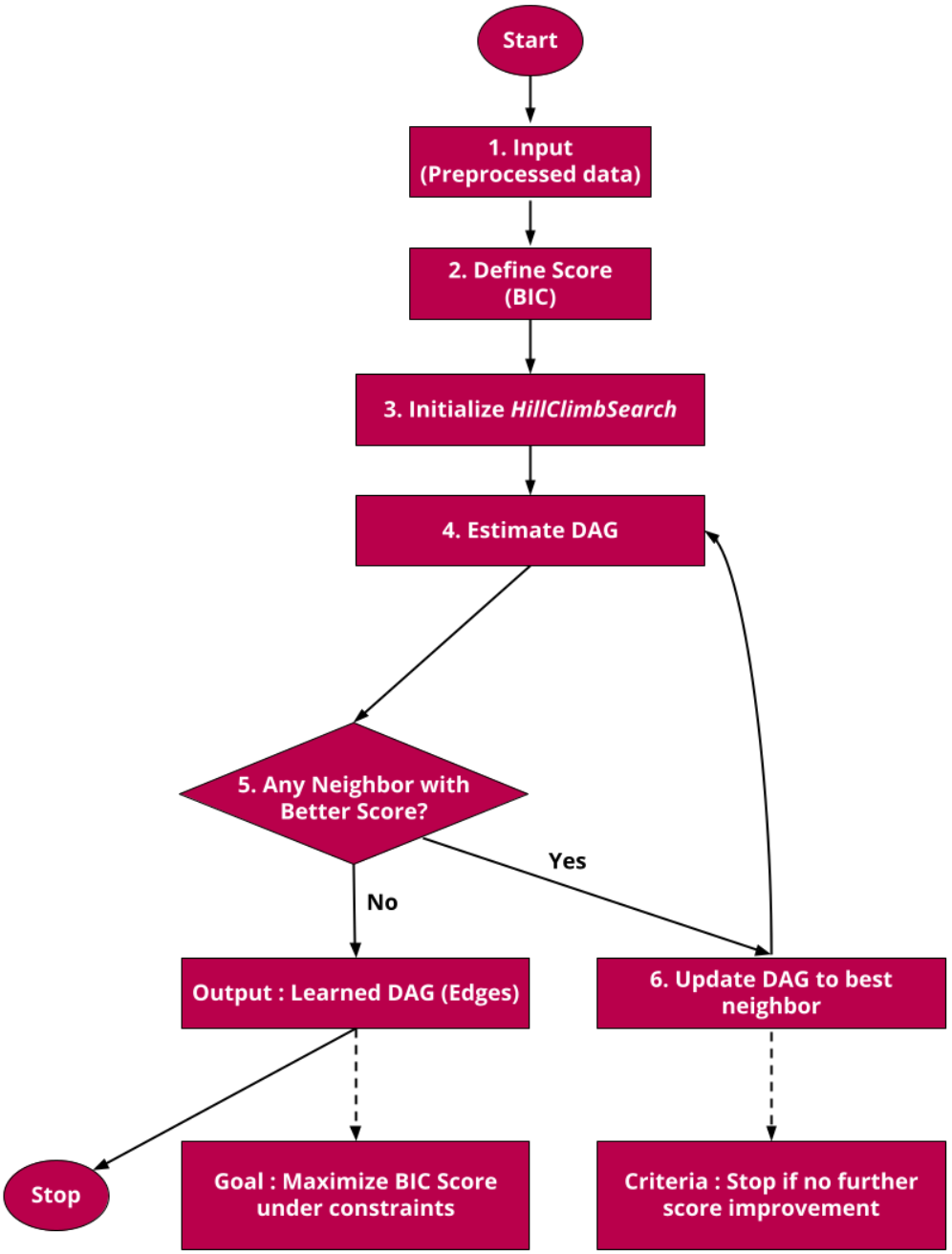



*Figure 12 : Flowchart du fonctionnement de HC*

## Algorithme LiNGAM (modèles causaux structurels)

DirectLiNGAM appartient à la famille des algorithmes basés sur la structure.
Il se base sur l'hypothèse que le bruit des données est indépendant et non-gaussien.

Son principe de fonctionnement repose sur l'identification des variables exogènes (*une variable qui ne dépend pas des autres variables*). Une fois une de ces dernières identifiées, elle est placée dans un ordre causal et son influence est retirée des autres variables. Ce processus permet d'estimer simultanément l'ordre causal des variables ainsi que les coefficients associés aux relations causales.

Contrairement aux algorithmes PC et HC, DirectLiNGAM produit un graphe entièrement orienté, permettant une interprétation directe du sens et de l'intensité des relations causales. En revanche, il est plus restrictif. En effet, il ne peut être appliqué qu'à des jeux de données composés uniquement de variables numériques continues.

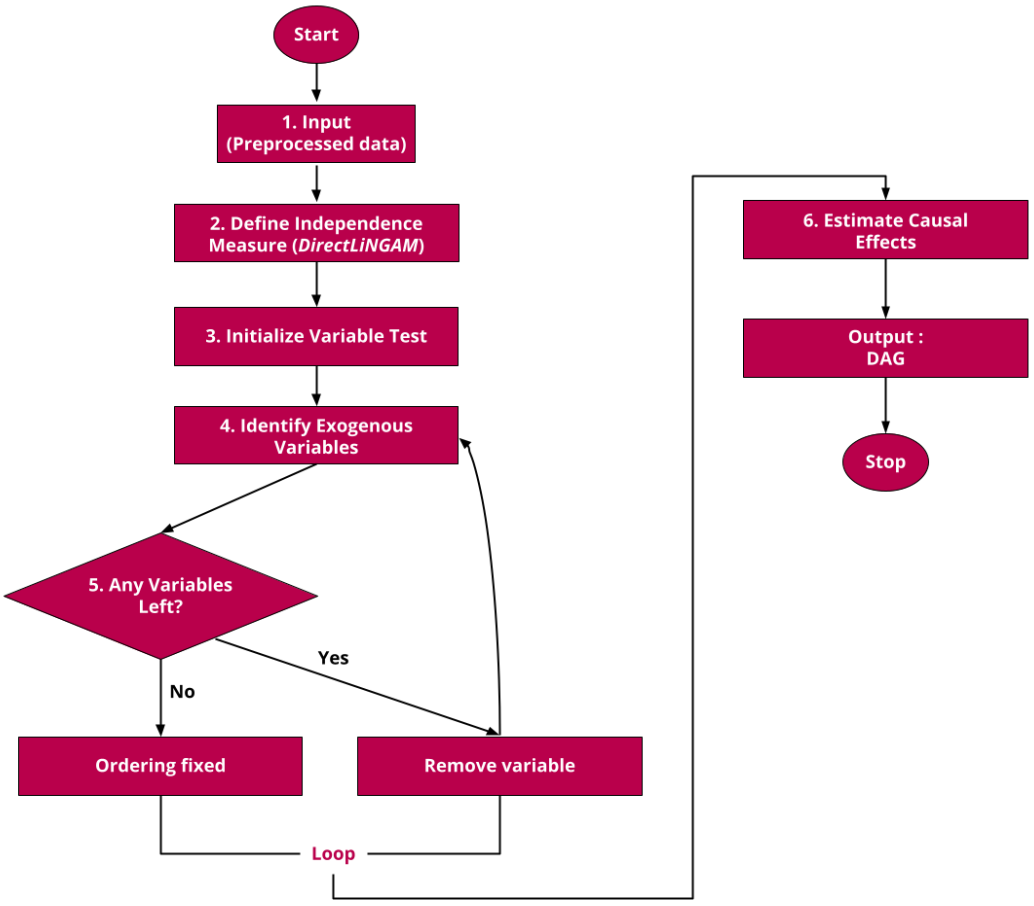



*Figure 13 : Flowchart du fonctionnement de LiNGAM*

# VI. Résultats

Dans cette section, nous analysons les graphes causaux inférés par les différents algorithmes pour chacun des jeux de données étudiés. Nous porterons notamment notre œil sur une lecture structurelle des graphes.

## 1er jeu de données, "*loan.csv*"

### Algorithme Peter-Clark

Le graphe causal obtenu à l'aide de l'algorithme PC met en évidence plusieurs relations cohérentes entre les variables décrivant le profil financier et l'historique de crédit des emprunteurs.

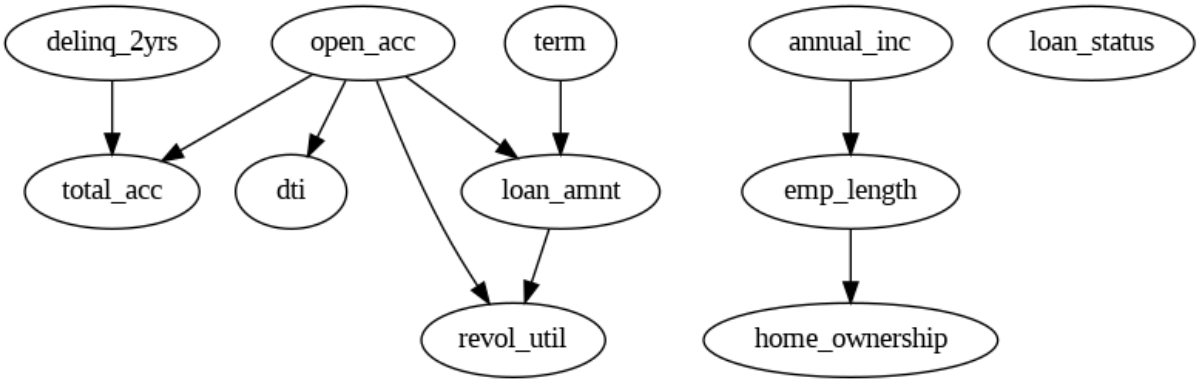



*Figure 14 : Graphe causal obtenu par PC sur loan.csv et ses variables sélectionnées*

On observe notamment un bloc fortement connecté autour de variables liées au crédit :

- le nombre de comptes ouverts (*open_acc*)
- le nombre total de comptes (*total_acc*)
- le ratio d'endettement (*dti*)
- le taux d'utilisation du crédit renouvelable (*revol_util*)
- le montant du prêt (*loan_amnt*)

Ces relations sont intuitivement logiques. En effet, un emprunteur disposant de nombreux comptes de crédit est susceptible d'avoir un historique plus complexe, influençant à la fois son endettement et son utilisation de crédit.

Cependant, la variable cible *loan_status* est totalement isolée, sans aucune relation directe avec les autres variables. Cela signifie que l'algorithme PC n'a identifié aucune dépendance conditionnelle entre *loan_status* et les autres variables sélectionnées.

Cette absence de connexions vient compromettre l'utilisation du graphe pour l'évaluation causale car la Markov Blanket de la variable cible sera vide. Ainsi, malgré une structure interne que nous trouvons cohérente, le jeu de données paraît peu exploitable.
*run time -->03:09s,30.40s/it*

## Algorithme Hill-Climbing

L'algorithme HC, appliqué avec le score BIC, produit un graphe causal avec une autre structure différente que précédemment mais partage tout de même des similitudes.

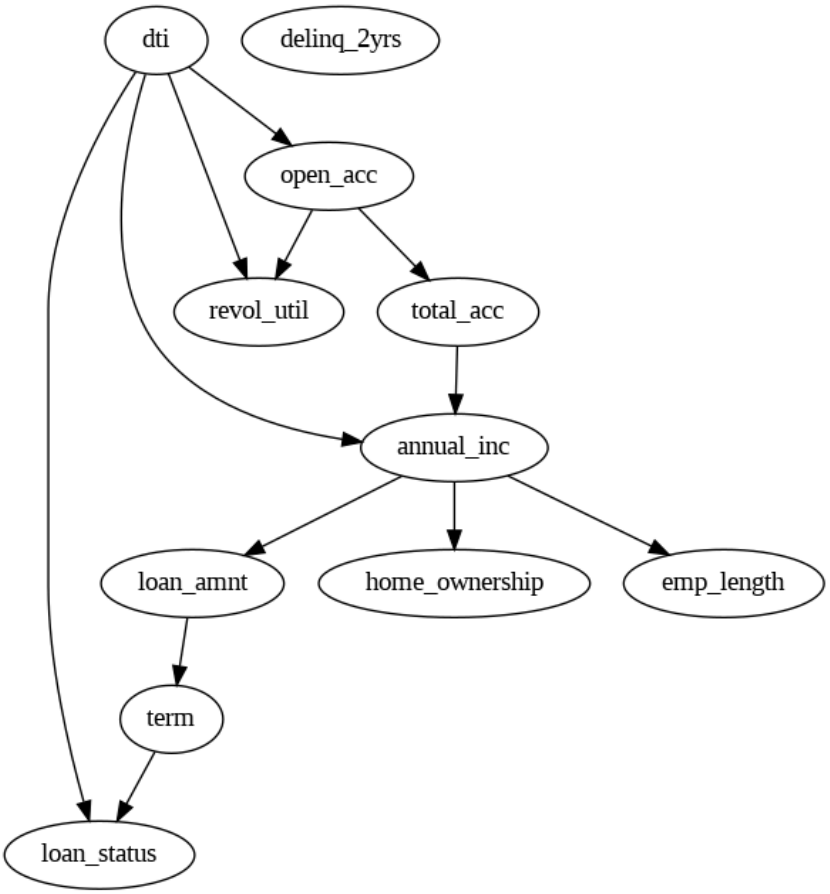



*Figure 15 : Graphe causal obtenu par HC sur loan.csv et ses variables sélectionnées*

Les relations causales identifiées concernent principalement les mêmes variables que celles mises en évidence par PC, notamment autour du revenu, de l'endettement et de l'historique de crédit.

Un peu de la même manière que la variable *loan_status*, cette dernière est faiblement connectée au reste du graphe.

De plus, cette convergence des résultats entre les deux approches suggère que la difficulté à établir des liens causaux directs avec la variable cible est davantage liée à la structure du jeu de données ou au choix des variables qu'au choix de l'algorithme.

*run time → <1s, 8.55it/s*

## 2ème jeu de données, "*GiveMeSomeCredit.csv*"

### Peter-Clark (avec les mêmes paramètres que précédemment)

Pour ce second jeu de données, nous appliquons l'algorithme PC avec les mêmes paramètres que ceux utilisés précédemment.
Ici, nous obtenons le graphe causal suivant :

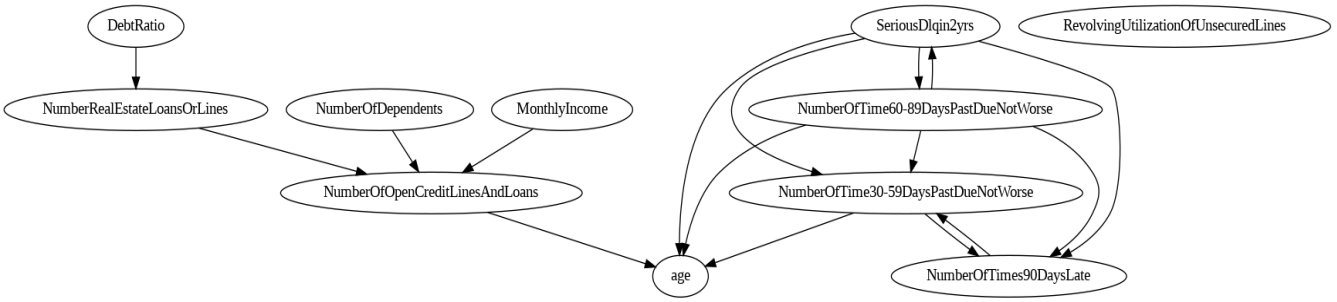



*Figure 16 : Graphe causal obtenu par PC sur GiveMeSomeCredit.csv*

Contrairement au premier jeu de données, la variable cible *SeriousDlqin2yrs* apparaît cette fois-ci connectée à plusieurs variables. De plus, les variables décrivant les retards de paiement passés (*NumberOfTimes30-59DaysPastDueNotWorse, NumberOfTime60-89DaysPastDueNotWorse et NumberOfTimes90DaysLate*) sont reliées à la variable cible.

D'un point de vue intuitif cette structure est cohérente car l'historique de retard constitue un facteur important dans la probabilité de rencontrer un défaut de paiement sérieux.
Ainsi, contrairement au jeu de données loan.csv, l'algorithme PC nous permettra ici d'en identifier une Markov Boundary non vide pour la variable cible.

Cependant, cette Markov Blanket reste tout de même inutilisable en effet les variables (*NumberOfTimes30-59DaysPastDueNotWorse, NumberOfTime60-89DaysPastDueNotWorse et NumberOfTimes90DaysLate)* sont trop corrélées à la variable cible, on en perds l'intérêt

de la découverte causale. Cela met en évidence l'importance de choisir les bons jeux de données.
*run time → 05:01s, 60.25s/it*

## Hill-Climbing (avec les mêmes paramètres que précédemment)

Ici de même, HC est appliqué avec les mêmes paramètres que précédemment en utilisant le critère BIC comme fonction de score.

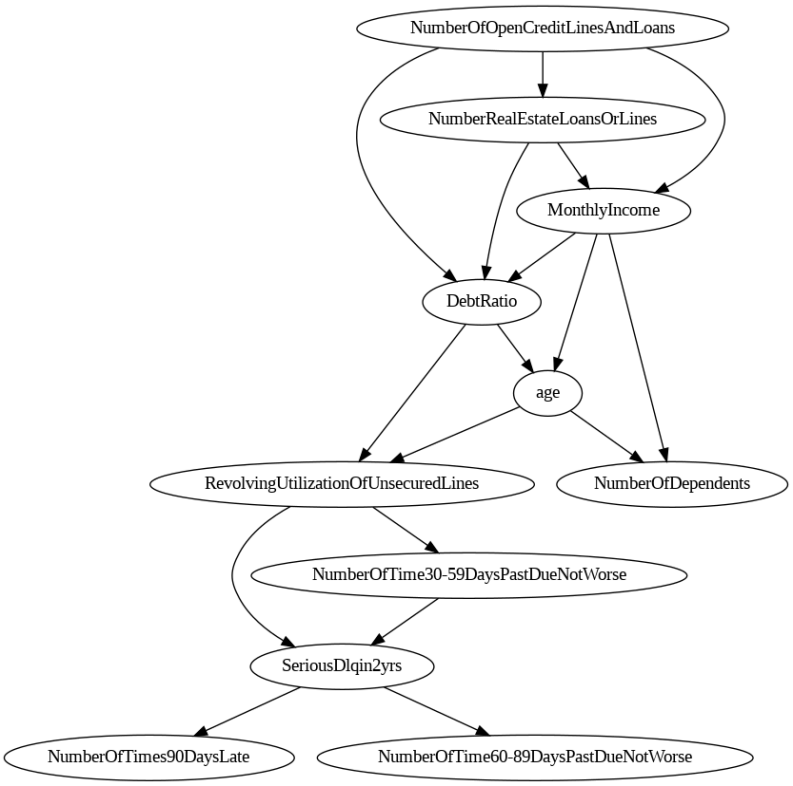



*Figure 17 : Graphe causal obtenu par HC sur GiveMeSomeCredit.csv*

Comme avec PC, nous retrouvons le même "bloc" autour de la variable cible, toujours liées aux retards de paiement (*NumberOfTimes30-59DaysPastDueNotWorse, NumberOfTime60-89DaysPastDueNotWorse et NumberOfTimes90DaysLate*).
*run time → <1s,26.89it/s*

## 3ème jeu de données : "*OnlineShoppers.csv*"

### Peter-Clark (avec les mêmes paramètres que précédemment) :

L'algorithme Peter-Clark est appliqué à OnlineShoppers.csv avec les mêmes paramètres que pour les jeux de données précédents. Voici ce que nous obtenons :

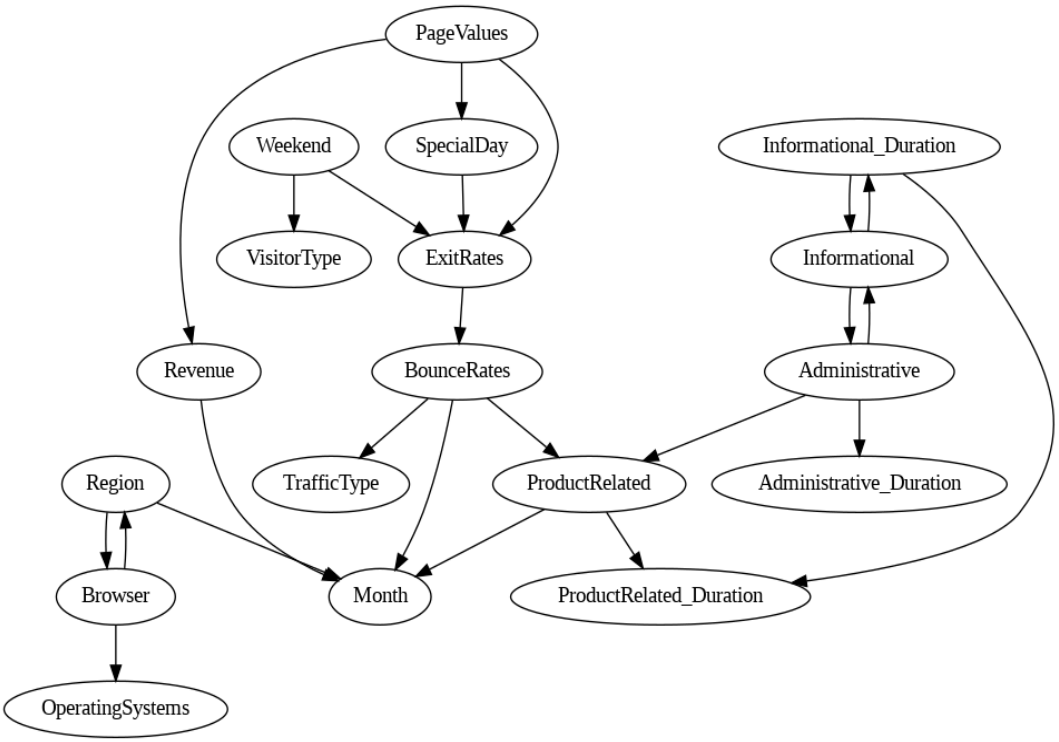



*Figure 18 : Graphe causal obtenu par PC sur OnlineShoppersIntention.csv*

Le graphe causal obtenu présente une structure relativement dense. On observe que la variable cible *Revenue* est directement reliée à plusieurs variables clés telles que la Markov Blanket [PageValues,BounceRates, ProductRelated, Month,Region] **5 features.**

Hill-Climbing (avec les mêmes paramètres que précédemment):

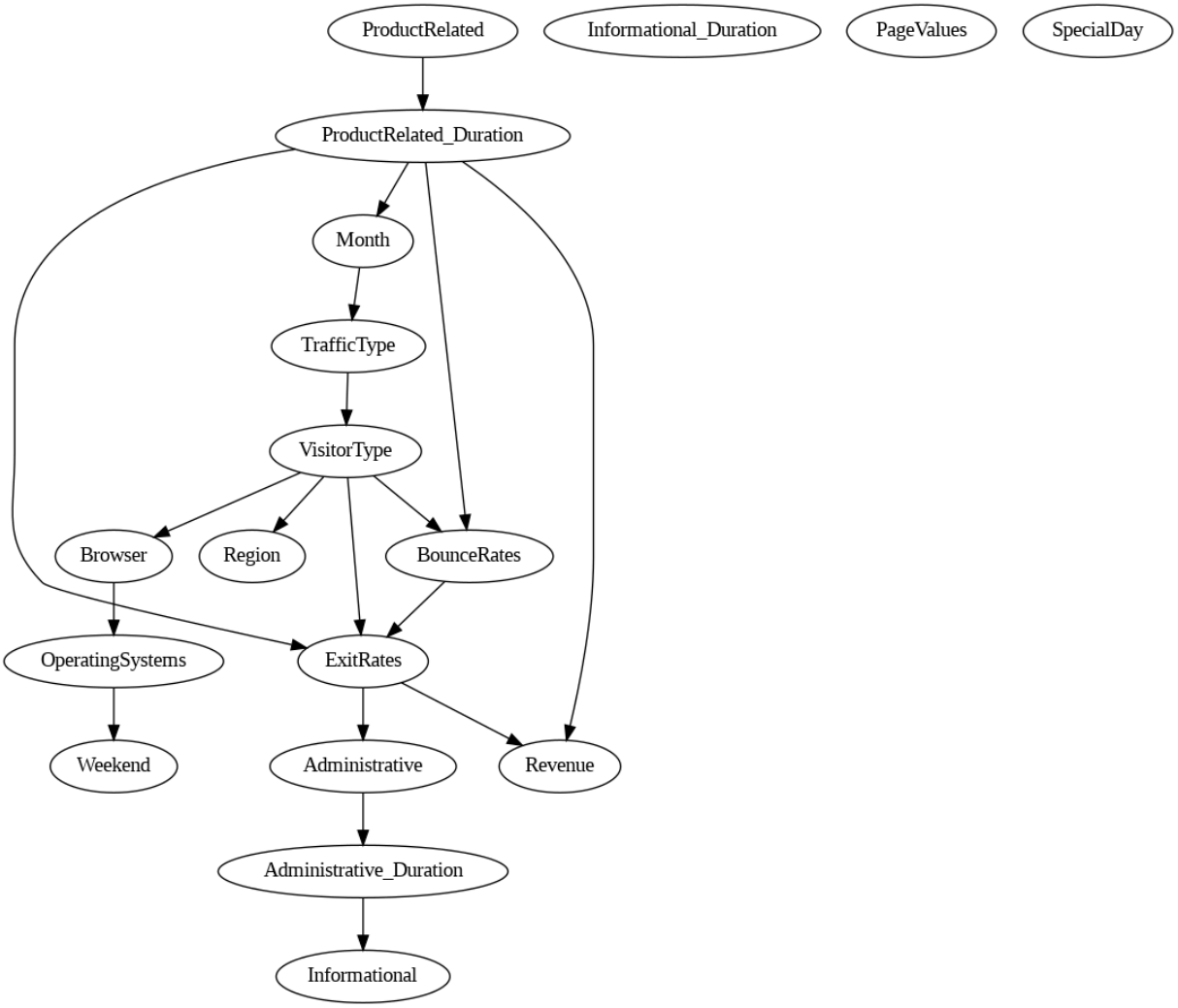



*Figure 19 : Graphe causal obtenu par HC sur OnlineShoppersIntention.csv*

De même on en extrait la Markov Blanket [ExitRates, ProductRelated_Duration]. Sachant qu'elle ne présente que 2 variables, on sait par avance que les résultats de puissance prédictive ne seront pas très bons. (**2 features**)

DirectLiNGAM

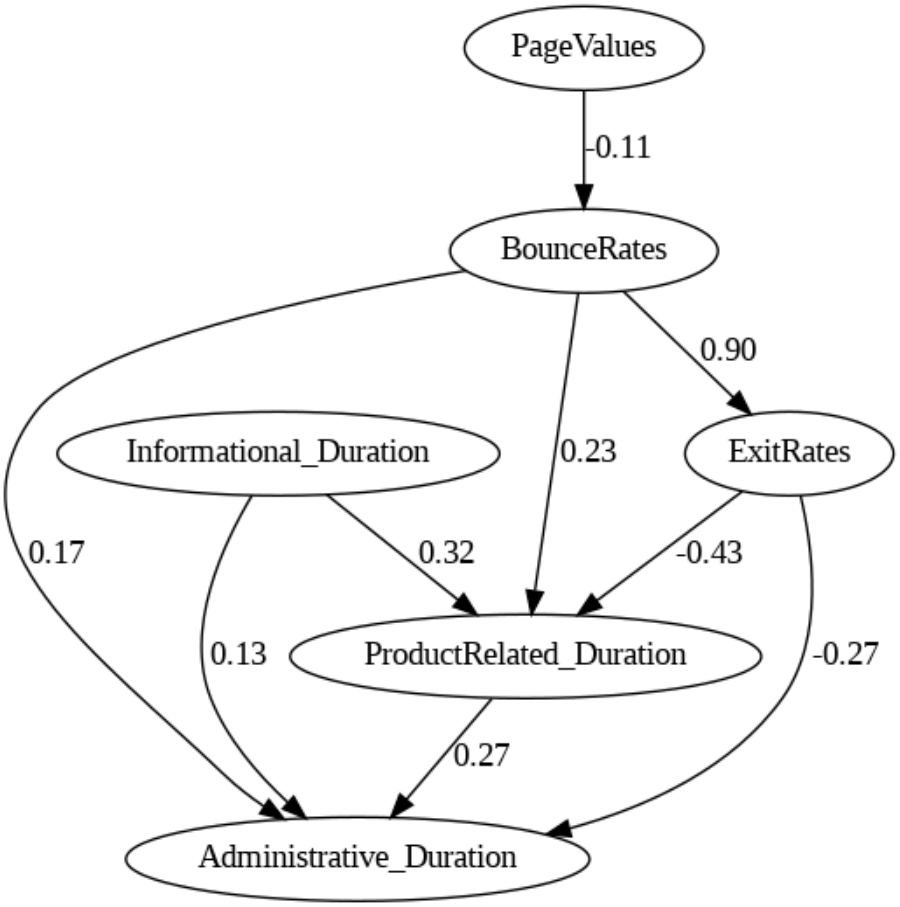
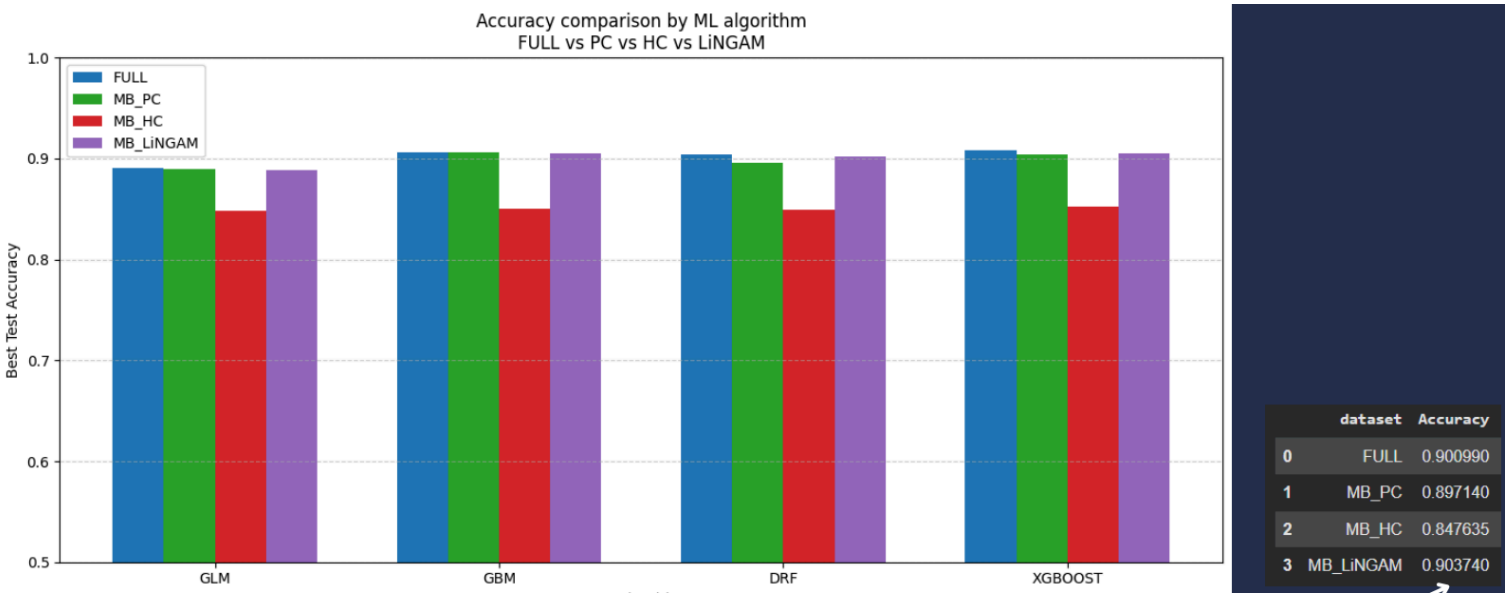



*Figure 20 : Graphe causal obtenu par DirectLiNGAM sur OnlineShoppersIntention.csv*

Ici, Lingam n'est utilisé que sur les variables continues du dataset. On récupère sa Markov Blanket : [SpecialDay,Weekend,BounceRates, PageValues, ProductRelated, Informational_Duration,ExitRates, Administrative] **8 features**

## Comparaison des algorithmes PC et HC sur OnlineShoppers.csv par puissance prédictive.

Premièrement, nous utilisons H20 autoML sur le dataset contenant toutes les variables. Nous utilisons autoML pour une raison simple, il faut avoir les algorithmes de prédiction les plus optimisés afin de pouvoir comparer équitablement les algorithmes de découvertes causales.
Les algorithmes utilisés pour la comparaison sont les suivants : **GLM, GBM, DRF, XGBOOST**.
Ces choix sont, pour le coup arbitraires, nous les avons choisis car ils comptent aujourd'hui parmi les algorithmes ML les plus répandus.




| | dataset | Accuracy |
|---|---|---|
| 0 | FULL | 0.900990 |
| 1 | MB_PC | 0.897140 |
| 2 | MB_HC | 0.847635 |
| 3 | MB_LiNGAM | 0.903740 |

*Figure 21 : Comparaison de précision entre la prédiction avec toutes les variables et les markov boundaries de chaque algorithme*

Les résultats sont très satisfaisants. En effet, la markov boundary du PC contient 5 variables contre 18 dans le full dataset et, nous obtenons quasiment des performances égales quant à la prédiction de la variable *Revenue*. Quant au LiNGAM contenant 8 variables, à des résultats quasi équivalent à ceux du FULL dataset et encore mieux, LiNGAM atteint la meilleur accuracy tout algorithme ML confondus.

Cependant, la puissance prédictive de la markov boundary du HC est plus faible. En effet, celle-ci ne contient que 2 variables, on pouvait s'y attendre.

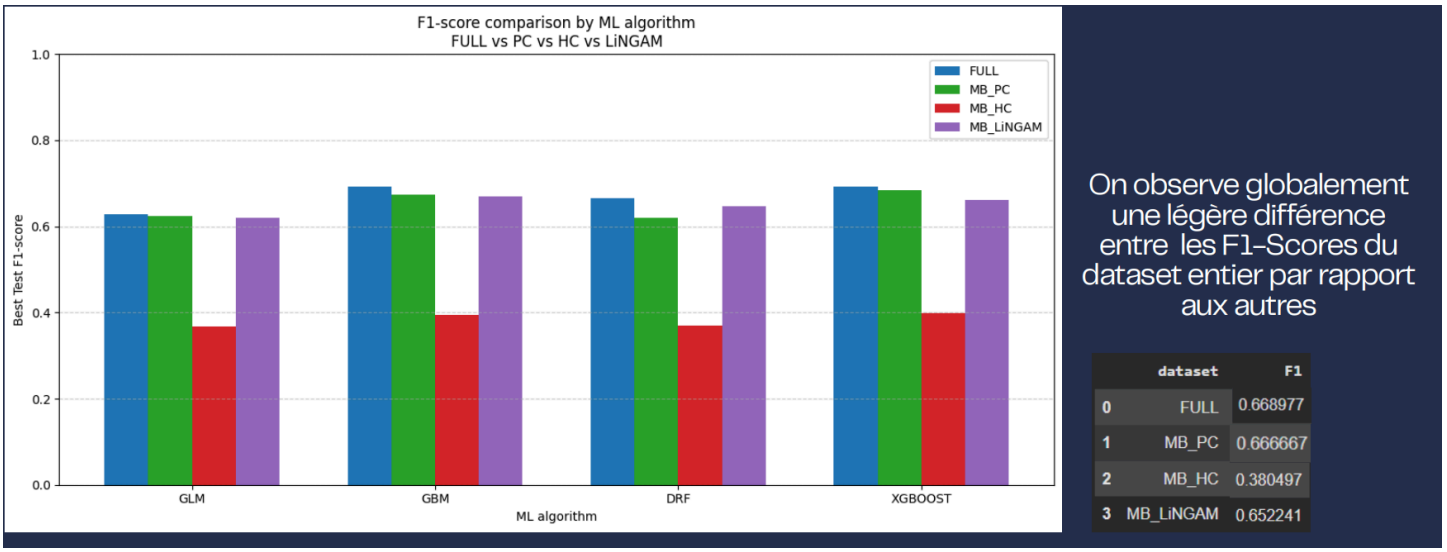



*Figure 22 : Comparaison des scores F1-scores entre la prédiction avec toutes les variables et les markov boundaries de chaque algorithmes*

Quant au F1-Score, les résultats sont tout aussi convaincants. Cette métrique est ici plus adaptée à notre target variable car le F1-Score prend en compte le Recall qui accorde plus d'importance à la détection d'instances positives, notamment dans un contexte de déséquilibre des classes.

# VII. Limites et perspectives

## Limites de l'étude

Ce travail repose sur l'analyse comparative de plusieurs algorithmes de découverte causale appartenant à des familles méthodologiques distinctes. Bien que cette approche permette de mettre en évidence des comportements différents selon la nature des données et des hypothèses, elle présente néanmoins plusieurs limites.

Tout d'abord, le choix des algorithmes étudiés implique des hypothèses fortes sur les données. Les algorithmes basés sur des contraintes (comme Peter-Clark ici), reposent sur des tests d'indépendance conditionnelle dont la complexité computationnelle augmente rapidement avec le nombre de variables. Cette caractéristique limite leur applicabilité à des jeux de données de grande dimension. À l'inverse, les algorithmes basés sur des scores (comme Hill-Climbing ici), peuvent converger vers des optima locaux ce qui rend les résultats sensibles aux choix initiaux et aux paramètres de pénalisation utilisés.

Par ailleurs, l'algorithme DirectLiNGAM impose des hypothèses encore plus restrictives, notamment la linéarité des relations causales et la non-gaussianité des bruits. Ces contraintes ont limité son application à certains jeux de données empêchant donc une comparaison avec les autres approches sur l'ensemble des datasets étudiés.

Une autre limite majeure de cette étude réside dans l'absence de graphe causal de référence. En l'absence de vérité terrain, l'évaluation des graphes causaux inférés ne peut être réalisée complètement. Le recours à la Markov Boundary et à la puissance prédictive associée constitue, selon nous, une approche pertinente, mais ne permet pas de valider formellement la validité structurelle des graphes obtenus.

Enfin, les résultats montrent que la performance des algorithmes dépend fortement de la nature des données considérées. Certains jeux de données, comme Loan.csv, se révèlent peu adaptés à une analyse causale centrée sur la variable cible, tandis que d'autres comme OnlineShoppersIntention.csv, favorisent l'émergence de relations causales interprétables.

## Perspectives

| Algorithme | Famille | Adapté lorsque | Limitations |
|---|---|---|---|
| PC Peter Clark | Basé sur les contraintes | • Nombre de données modéré<br>• Données de types mixtes | • Complexité computationnelle élevée |
| HC Hill-Climbing | Basé sur le score | • Bon compromis entre performance et temps de calcul<br>• Données mixtes ou discrétisées | • Risque de convergence vers des optima locaux |
| L LiNGAM | Basé sur la structure | • Variables principalement continues<br>• intérêt pour l'intensité des effets causaux | • Hypothèse de non-gaussianité du bruit requise |

*Figure 23 : Choix des algorithmes de découvertes causales basé selon notre étude*

Malgré ces limites, ce travail ouvre des perspectives intéressantes quant à l'intégration de la découverte causale dans des pipelines de machine learning appliqués à des jeux de données complexes et de grande dimension.
Dans un contexte où l'entraînement de modèles de machine learning devient de plus en plus coûteux, la découverte causale apparaît comme un levier potentiel de réduction de la complexité. En identifiant des sous-ensembles de variables pertinentes, tels que la Markov Boundary de la variable cible, il devient possible de réduire significativement le nombre de variables utilisées lors de l'entraînement, tout en maintenant, voire en améliorant, les performances prédictives.

Cette approche permettrait non seulement de diminuer les coûts computationnels, mais également de renforcer l'interprétabilité des modèles, en concentrant l'analyse sur des variables ayant une justification causale claire. À plus long terme, ces résultats ouvrent la voie à des pipelines hybrides combinant découverte causale et apprentissage supervisé, dans lesquels la découverte causale agirait comme un filtre en amont de modèles prédictifs plus complexes.

# VIII. Annexes

## Jeux de données :

Lending Club Loan Data

Give Me Some Credit | Kaggle

Online Shoppers Purchasing Intention Dataset - UCI Machine Learning Repository

## Sources :

Hillclimb-Causal Inference: a data-driven approach to identify causal pathways among parental behaviors, genetic risk, and externalizing behaviors in children | Journal of the American Medical Informatics Association | Oxford Academic

Causal Discovery with Mixed Data using pgmpy | by Ankur Ankan | Medium

DirectLiNGAM — LiNGAM 1.11.0 documentation

The History and Development of Search Methods for Causal Structure by FREDERICK EBERHARDT

Causal discovery by DARIA BYSTROVA

Evaluation and Comparison of Causal Discovery Algorithms Bachelor´s Thesis by Hui Gong Department of Informatics by HUI GONG

Empirical Analysis of Filter Feature Selection Criteria on Financial Datasets

A Comprehensive Review of Causal inference in Banking, Finance, and Insurance

# IX. Table des figures