Revue internationale de mathématiques appliquées, d'électronique et d'informatique 8(4) : 109-114, 2020

IJAMEC
e-ISSN : 2147-8228

REVUE INTERNATIONALE DE MATHÉMATIQUES APPLIQUÉES, ÉLECTRONIQUE ET INFORMATIQUE

www.dergipark.org.tr/ijamec

Accès libre international

Volume 08
Numéro 04

Décembre 2020

Article de recherche

Analyse et détection des survivants du Titanic à l'aide de modèles linéaires généralisés et d'un algorithme d'arbre de décision

Burcu Durmuş [a], ***Öznur İşçi Güneri*** [a, *]

[a] *Université Mugla Sitki Kocman, Unité rectorale, Campus Kotekli, Mugla, Turquie*
[b] *Université Mugla Sitki Kocman, Faculté des sciences, Département de statistique, Campus Kotekli, Mugla, Turquie*

INFORMATIONS SUR L'ARTICLE

Historique de l'article :
Reçu le 25 août 2020
Accepté le 9 octobre 2020

Mots-clés :
Arbre de décision,
Modèles linéaires généralisés,
Régression logit,
Régression probit,
Arbre aléatoire.

RÉSUMÉ

Cet article vise à étudier les facteurs qui influencent la survie dans les accidents majeurs légendaires d'aujourd'hui à l'aide de différentes méthodes. L'analyse a pour objectif de trouver la méthode qui détermine le mieux la survie. À cette fin, des modèles logit et probit issus de modèles linéaires généralisés et un algorithme d'arbre aléatoire issu de méthodes d'arbre de décision ont été utilisés. L'étude a été menée en deux étapes. Tout d'abord, l'analyse réalisée à l'aide de modèles linéaires généralisés a permis de déterminer les variables qui ne contribuaient pas de manière significative au modèle. La précision de la classification s'est avérée être de 79,89 % pour le modèle logit et de 79,04 % pour le modèle probit. Dans un deuxième temps, l'analyse de classification a été réalisée à l'aide d'arbres de décision aléatoires. La précision de la classification a été déterminée à 77,21 %. En outre, selon les résultats obtenus à partir des modèles linéaires généralisés, l'analyse de classification a été répétée en supprimant les données qui n'apportaient aucune contribution significative au modèle. Le taux de classification a augmenté de 4,36 % pour atteindre 81,57 %. Au final, il a été déterminé que l'analyse par arbre de décision réalisée avec les variables extraites du modèle donnait de meilleurs résultats que l'analyse réalisée avec les variables d'origine. Ces résultats sont considérés comme utiles pour les chercheurs travaillant sur l'analyse de classification. En outre, les résultats peuvent être utilisés à des fins telles que le prétraitement et le nettoyage des données.

1. Introduction

Le Titanic est un paquebot de croisière mondialement connu qui a coulé lors de son premier voyage dans l'Atlantique Nord [1]. La littérature regorge de spéculations sur le naufrage légendaire du Titanic, et les recherches à ce sujet se poursuivent encore aujourd'hui [2-3]. Au fil des ans, un ensemble de données contenant des informations sur les survivants ainsi que sur les passagers et les membres d'équipage décédés a été créé [4]. Cet ensemble de données est accessible au public sur Kaggle.com [5].

Lorsque l'on examine la littérature, il apparaît clairement que les données du Titanic ont été examinées à des fins différentes ces dernières années. Dans l'étude de Barhoom et al., la prédiction des survivants a été déterminée par des réseaux neuronaux artificiels. L'algorithme a atteint une précision de 99,28 % [6]. Singh et al. ont étudié les données du Titanic à l'aide de la régression logistique, de l'arbre de décision, de l'arbre de décision avec hypertuning, des k plus proches voisins et des machines à vecteurs de support. À la fin de l'étude, ils ont obtenu la meilleure estimation avec les arbres de décision, soit 93,6 % [7]. Kakde et al., quant à eux, ont effectué l'analyse à l'aide de méthodes de régression logistique, d'arbre de décision, de forêt aléatoire et de machines à vecteurs de support en utilisant le nettoyage des données. Ils ont suggéré que, dans l'idéal, la régression logistique et la machine à vecteurs de support offrent un bon niveau de précision lorsqu'il s'agit du problème de classification [8]. Dans une autre étude, Kshirsagar et al. ont montré que les survivants du Titanic pouvaient être prédits par régression logistique avec une précision de 95 % [4].

Avec le développement de la technologie, la collecte et le stockage des données sont devenus assez faciles. Il est donc devenu plus important de découvrir de nouvelles méthodes d'analyse des données. De nombreux progrès ont été réalisés dans ce domaine ces dernières années. Des mesures importantes ont été prises, en particulier dans le domaine de l'exploration de données. De nombreux nouveaux algorithmes ont été introduits et les algorithmes existants ont également été améliorés. À la suite de ces développements, l'objectif est désormais d'obtenir des résultats différents et nouveaux en analysant les données à l'aide de différentes méthodes, comme c'est le cas avec

* Auteur correspondant. Adresse e-mail : *oznur.isci@mu.edu.tr*
DOI : 10.18100/ijamec.785297

les chercheurs travaillant sur les données du Titanic.

Dans cette étude, contrairement à la littérature, les données du Titanic ont été analysées à l'aide de l'algorithme Random Tree et de modèles linéaires généralisés. L'objectif principal de l'étude est de déterminer les caractéristiques des survivants du naufrage du Titanic à l'aide de différentes méthodes. Dans cette optique, les modèles de régression logit et probit, qui sont des modèles linéaires généralisés, ont été examinés dans un premier temps. À ce stade, un test de signification a d'abord été appliqué aux données et les variables qui contribuaient de manière significative au modèle ont été incluses dans l'analyse. Dans un deuxième temps, l'analyse a été réalisée à l'aide de l'algorithme Random Tree, qui est l'algorithme d'apprentissage par arbre de décision. Afin d'améliorer la réussite du modèle, l'analyse de classification Random Tree a été répétée avec les variables qui contribuaient de manière significative au modèle. L'étude a été complétée par une comparaison des résultats.

2. Méthodes et matériel

Dans cette étude, les modèles logit et probit de la famille généralisée des modèles linéaires et l'arbre de décision des méthodes d'exploration de données sont examinés.

2.1. Ensemble de données Titanic

L'ensemble de données contient les variables indiquées dans le tableau 1. Cependant, l'analyse logit et probit binaire a permis de déterminer que certaines de ces variables (sibsp, parch, embarked) n'apportaient pas de contribution significative au modèle. Elles ont donc été supprimées de l'ensemble de données. Les variables restantes ont été incluses dans le modèle logit et probit en tant que données catégorielles. Une analyse statistique descriptive a été effectuée avec les variables du programme SPSS 22.0. Dans la suite de l'étude, des analyses logistiques binaires et probit binaires ont été réalisées avec le programme Stata 11.0 et une analyse de classification par arbre de décision a été effectuée. L'analyse par arbre de décision a été réalisée à la fois avec les variables restantes de l'ensemble de données et avec la version originale de l'ensemble de données.

Tableau 1. Variables de l'ensemble de données

Variables	Définition
survivant	non : 0, oui : 1
pclass	classe de passager (1, 2, 3)
sex	féminin, masculin
âge	âge
nombre de frères et sœurs	nombre de frères et sœurs ou conjoints à bord
parch	nombre de parents ou d'enfants à bord
tarif	tarif passager
embarqué	port d'embarquement

Des analyses de régression logit et probit binaires ont été effectuées en déterminant des variables indicatrices. Variables indicatrices : pclass1, sexe masculin, âge0 (enfants), tarif0.

2.2. Modèles linéaires généralisés

Les modèles linéaires généralisés sont obtenus en étendant les modèles linéaires en raison des distorsions liées aux hypothèses [9]. Dans de nombreux domaines, ces modèles sont utilisés si les données sont catégorielles ou discontinues [10]. Les modèles linéaires généralisés se composent d'une composante aléatoire, d'une composante systématique et d'une fonction de liaison. La fonction de liaison détermine le nom du modèle utilisé. Si une liaison logit est utilisée, le nom du modèle est appelé modèle de régression logit [11]. Dans cette étude, les modèles logit et probit sont examinés.

2.2.1. Régression logit

Si le lien canonique utilisé dans les modèles linéaires généralisés est logit, le modèle est une régression logit [12]. La régression logistique est indépendante des variables lorsque la variable dépendante est catégorielle, binaire ou multiple. Dans la régression logit, il n'y a pas d'hypothèse de normalité et de continuité [13]. On peut donc dire qu'elle est plus flexible que les modèles linéaires.

Le modèle logit est dérivé de la fonction de distribution cumulative donnée par l'équation 1 [14].

$$P_{i=} E(Y = x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_i)}} \quad (1)$$

Dans ce modèle,*Pi* fournit des informations sur l'argument*xi* tandis que le premier individu exprime la probabilité de faire un choix particulier [15]. Ainsi, *Pi* prend également des valeurs comprises entre « 0 » et « 1 » [16]. Lorsque le taux de réalisation d'un événement est divisé par le taux de non-réalisation de l'événement, on obtient le rapport de cotes [17].

Il devient linéaire lorsque l'on prend le logarithme du rapport de cotes. Dans ce cas, le modèle est appelé logit et l'équation donnée par l'équation 2 est appelée fonction de liaison logit.

$$L_i = log\left(\frac{P_i}{1 - P_i}\right) \quad (2)$$

Il existe 3 méthodes de base dans l'analyse de régression logistique :

- Binaire
- Ordinal
- Nominale

2.2.2. Régression probit

Tout comme le modèle logit, ce modèle garantit que les probabilités restent comprises entre 0 et 1. Le modèle probit suppose que la variable dépendante est distribuée normalement. Par conséquent, le graphique du modèle logit est plus large que celui du modèle probit (Fig. 1). Les modèles logit et probit peuvent être comparés à l'aide d'un coefficient proposé par Amemiya [18].

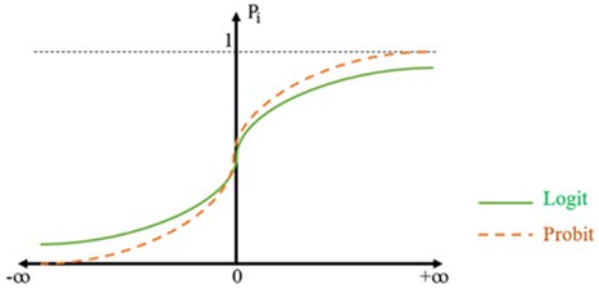

Figure 1. Distributions logit et probit

Lorsque la distribution des erreurs est la distribution cumulative normale standard, la fonction de liaison probit est utilisée

et le modèle est appelé modèle probit [19]. La fonction de liaison probit est définie par l'équation 3.

$$Z = \varphi^{-1}(\mu) = \sum_{k=1}^{K} b_k x_k \quad (3)$$

Ici, φ^{-1} désigne l'inverse de la distribution normale standard, b_k est l'estimation des coefficients et x_k est les variables explicatives. u pour montrer l'erreur pour chaque œil ; la fonction de distribution cumulative standard est donnée par
Équation 4.

$$\varphi(Z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} exp\left(\frac{-u^2}{2}\right) du \quad (4)$$

2.3. Algorithme d'arbre aléatoire

En raison de ses nombreux avantages, l'apprentissage par arbre de décision est souvent utilisé dans les études d'exploration de données [20]. Une structure arborescente est créée dans l'apprentissage par arbre de décision. L'arbre commence à partir du nœud racine et, à partir de là, la structure est divisée en nœuds internes. Le nœud racine peut être considéré comme la caractéristique la plus déterminante des différences entre les données. Il est divisé en nœuds internes après une série d'opérations appliquées à l'ensemble de données. Chaque nœud peut être divisé en plusieurs nœuds internes. Le nœud feuille est atteint en contrôlant tous les nœuds internes. Le nœud feuille est l'endroit où la décision est prise. Chaque transition entre les nœuds dépend d'une condition. La condition mentionnée ici est la théorie sur laquelle repose l'algorithme choisi [21-22]. Les arbres de décision présentent de nombreux avantages, notamment un faible coût de calcul et une grande facilité de compréhension. C'est pourquoi, comme mentionné au début, ils sont privilégiés dans de nombreuses études d'exploration de données et en particulier dans les études de classification [23-24].

L'algorithme d'arbre aléatoire est une méthode dans laquelle plusieurs arbres de décision sont créés [25]. Étapes de l'algorithme :

- La caractéristique qui fournit la meilleure classification est sélectionnée et le nœud de départ est créé.
- Un ensemble d'apprentissage est formé à partir d'une partie de l'ensemble de données. Les données restantes constituent l'ensemble de test.
- Des arbres sont créés avec le nombre de variables à utiliser dans chaque nœud et le nombre d'arbres dans N. Les variables sont sélectionnées de manière aléatoire à chaque nœud.
- Lorsque N arbres sont produits, le modèle est terminé et la classe du nouveau membre est estimée [25-26].

2.4. Matrice de confusion

La matrice de confusion est un outil d'analyse qui explique les observations correctement classées et les observations incorrectement classées. La matrice de confusion est l'état d'un ensemble de données et le nombre de prédictions correctes et incorrectes de notre modèle de classification converti en tableau. La forme générale de la matrice de confusion est présentée dans le tableau 2.

Tableau 2. Matrice de confusion générale

Classe réelle	Classe prédite	
	Positifs	Négatifs
Positifs	TP (vrais positifs)	FN (faux négatifs)
Négatifs	FP (faux positifs)	TN (vrais négatifs)

3. Résultats

3.1. Statistiques descriptives

Lorsque l'on examine la relation entre les variables « survie » et « sexe » dans le tableau 3, on constate que 359 personnes sont décédées et 93 ont survécu, 64 femmes sont décédées et 195 ont survécu.

Autrement dit, sur les 423 personnes décédées, 359 étaient des hommes et 64 des femmes. De même, sur les 288 survivants de l'accident, 93 sont des hommes et 195 des femmes.

Tableau 3. Relation entre les variables « survie » et « sexe »

Survivants	**Sexe**		Total
	0	1	
0	359	64	423
1	93	195	288
Total	452	259	711

En outre, on peut dire qu'il existe une concordance significative entre les variables « survivants » et « sexe » dans le tableau 4.

Tableau 4. Concordance entre les variables « survivants » et « sexe »

	Valeur	Asymp. Sig. (bilatéral)	Sig. exacte (2 côtés)	Sig. exacte (1-côté)
Chi-carré de Pearson carré	204,540	0,00	0,00	
Probabilité Ratio	210,756	0		
Exact de Fisher de Fisher				

Lorsque l'on examine la relation entre les variables « survie » et « tarif » dans le tableau 5, on constate que 286 des 429 personnes ayant payé un tarif bas sont décédées et 143 ont survécu ; sur les 282 personnes ayant payé un tarif élevé, 137 ont survécu et 145 sont décédées.

Tableau 5. Relation entre les variables « survie » et « tarif »

Survivants	**Tarif**		Total
	0	1	
0	286	137	423
1	143	145	288
Total	429	282	711

On peut dire qu'il existe une relation statistiquement significative ($p=0,00<0,05$) entre les variables de survie et de tarif dans le tableau 6.

Tableau 6. Harmonie entre les variables de survie et de tarif

	Valeur	Asymp. Sig. (bilatéral)	Sig. exacte (2 côté)	Sig. exacte (1-côté)
Chi-carré de Pearson carré	23,093	0,00	0,00	
Probabilité Ratio	23,029	0		
Exact de Fisher Test				

3.2. Résultats de la régression logit

En examinant le tableau 7, on peut dire que le modèle prédit est significatif à un niveau d'erreur de 5 %, puisque p = 0,00 < 0,05.

Lorsque l'on examine les valeurs de signification des variables, on constate que toutes les variables ont une contribution significative au modèle (celles qui n'avaient pas de contribution significative ont déjà été supprimées).

Tableau 7. Résultats du modèle de régression logit binaire

		Nombre d'observations LR Chi2 (8) Prob > chi2 Pseudo R2		= = = =	711 317,89 0,000 0,3312	
Survivants	Rapport de cotes	Erreur standard	z	P > |z|	[Intervalle de confiance à 95 %]	
pclass2	0,253	0,074	-4,67	0,000	0,14	0,45
pclass3	0,066	0,021	-8,46	0,000	0,03	0,12
Femmes	12,90	2,732	12,07	0,000	8,51	19,5
âge1	0,217	0,099	-3,32	0,000	0,08	0,53
âge2	0,225	0,082	-4,09	0,000	0,11	0,46
âge3	0,204	0,075	-4,29	0,000	0,09	0,42
âge4	0,090	0,044	-4,92	0,000	0,03	0,23
tarif1	0,610	0,150	-2,01	0,045	0,37	0,98
_cons	6,561	3,160	3,91	0,000	2,55	16,8

On peut dire qu'il existe une relation statistiquement significative (p = 0,00 < 0,05) entre les variables de survie et de tarif dans le tableau 5.

Le rapport de cotes est interprété par inversion. Les commentaires sur les rapports de cotes sont les suivants :

- Les passagers de 2e classe ont 6,55 fois plus de chances de survivre que ceux de 1re classe.
- Les passagers de 1re classe ont 4 fois plus de chances de survivre que ceux de 3e classe.
- Les hommes ont 12,80 fois plus de chances de survivre que les femmes.
- Les enfants ont 4,76 fois plus de chances de survivre que les personnes âgées de 1 an.
- Les enfants ont 4,54 fois plus de chances de survivre que les personnes âgées de 2 ans.
- Les enfants ont 5 fois plus de chances de survivre que les personnes âgées de 3 ans.
- Les enfants ont 11,11 fois plus de chances de survivre que les personnes âgées de 4 ans.
- Les petits contribuables ont 1,63 fois plus de chances de survivre que les gros contribuables.

Les valeurs de probabilité pour les données sont calculées à l'aide de l'équation 4. Quelques exemples de valeurs de probabilité sont donnés avec les équations 5 et 6.

$$p = 1{,}88 - 1{,}37 * class2 - 2{,}70 * class3 + 2{,}55 * famale - 1{,}52 * age1 - 1{,}48 age2 - 1{,}58 * age3 - 2{,}40 * age4 - 0{,}49 fare1 \quad (4)$$

La probabilité de survie pour une personne de 1re classe, femme, enfant, tarif élevé est $p = 0{,}98$.

La probabilité de survie d'une personne de 3e classe, de sexe masculin, âgée de 2 ans et payant un tarif élevé est d'$p = 0{,}09$.

L'effet marginal est l'effet qu'un petit changement dans la variable indépendante aura sur la variable dépendante.

Pour le modèle logit présenté dans le tableau 8, alors que l'effet des autres variables est fixe, une augmentation de 1 unité de la variable âge-1 diminue la survie de 0,22 unité en moyenne. Ce résultat est conforme aux résultats du rapport de cotes.

Tableau 8. Effet marginal du modèle de régression logit binaire

Effets marginaux moyens Modèle VCE : OIM Expression : Pr(Survivant), prédire () dy/dx par rapport à : âge1						
	dy/dx	Erreur standard	z	P > |z|	[Intervalle de confiance à 95 %]	
âge1	-0,22	0,065	-3,40	0,001	-0,34	-0,09

Les résultats de classification pour le modèle de régression logit sont présentés dans le tableau 9. Le modèle affiche une précision de classification de 79,89 %.

Tableau 9. Résultats de classification du modèle de régression logit binaire

Valeur réelle	Valeur prédite		Total réel
	0	1	
0	207	62	269
1	81	361	442
Total	288	423	711
Précision de la classification : 79,89 %			

3.3. Résultats de la régression probit

Selon le tableau 10, le modèle est significatif puisque p = 0,00 < 0,05. Au moins une variable a un effet sur le modèle. Les coefficients sont également importants, à l'exception de la variable fare1.

Tableau 10. Résultats du modèle de régression logit binaire

		Nombre d' observations LR Chi2 (8) Prob > chi2 Pseudo R2		= = = =	711 314,42 0,000 0,3276	
Survivants	Coef.	Err. Err.	z	P > |z|	[Intervalle de confiance à 95 % Intervalle]	
pclass2	-0,78	0,16	-4,61	0,000	-1,11	-0,45
pclass3	-1,50	0,17	-8,66	0,000	-1,84	-1,16
Femmes	1,49	0,11	12,64	0,000	1,26	1,72
âge1	-0,81	0,25	-3,16	0,000	-1,32	-0,30
âge2	-0,77	0,19	-3,88	0,000	-1,16	-0,38
âge3	-0,82	0,20	-4,07	0,000	-1,22	-0,42
âge 4	-1,31	0,27	-4,79	0,000	-1,85	-0,77
tarif1	-0,25	0,13	-1,84	0,066	-0,52	0,01
_cons	0,96	0,26	3,64	0,000	0,44	1,49

Les résultats de classification pour le modèle de régression probit sont présentés dans le tableau 11. Le modèle affiche une précision de classification de 79,04 %.

Tableau 11. Résultats de classification du modèle de régression probit binaire

Valeur réelle	Valeur prédite		Total réel
	0	1	
0	201	62	263
1	87	361	448
Total	288	423	711
Précision de la classification : 79,04 %			

Les résultats obtenus avec le modèle probit sont parallèles à ceux obtenus avec le modèle logit.

3.4. Résultats de l'algorithme Random Tree

L'algorithme Random Tree est examiné pour les modèles logit et probit et les variables qui contribuent de manière significative au modèle parmi les variables d'origine. Une section de l'arbre de décision obtenu par l'algorithme Random Forest est présentée dans le tableau 12. En examinant la structure de l'arbre, on peut voir comment les variables affectent la survie.

Tableau 12. Une section de l'arbre de décision

```
sexe = masculin
| âge = 0
| | pclass = 1 : 1 (3/0)
| | pclass = 2 : 1 (9/0)
| | classe = 3
| | | tarif = 0 : 1 (9/1)
| | | tarif = 1 : 0 (15/1)
| | | tarif = 2 : 0 (0/0)
| âge = 1
| | classe = 1 : 0 (2/1)
| | classe = 2 : 0 (6/0)
| | classe = 3
| | | tarif = 0 : 0 (22/2)
| | | tarif = 1 : 0 (4/0)
| | | tarif = 2 : 0 (0/0)
| âge = 2
| | classe = 1
| | | tarif = 0 : 0 (0/0)

| | | tarif = 1 : 1 (17/8)
```

Les résultats de classification de l'algorithme sont présentés dans le tableau 13. Le résultat de classification est plus satisfaisant, avec environ 2,5 % de plus que le résultat obtenu avec les modèles logit et probit.

Tableau 13. Matrice de confusion de l'ensemble de données

Valeur réelle	Valeur prédite		Total réel
	0	1	
0	383	40	423
1	91	197	288
Total	474	237	711
Précision de la classification : 81,57 %			

Une section de l'arbre de décision de l'ensemble de données original est présentée dans le tableau 14. Toutes les variables de l'ensemble de données sont incluses dans l'arbre de décision. On constate également que cette structure arborescente commence par la variable sexe, comme dans le tableau 12.

Tableau 14. Extrait de l'arbre de décision (pour l'ensemble de données d'origine)

```
sexe = masculin
| tarif = 0
| | embarqué = C
| | | âge < 29,5
| | | | âge < 5,5 : 1 (1/0)
| | | | âge >= 5,5
| | | | | nombre de frères et sœurs = 0
| | | | | | classe = 1 : 0 (0/0)
| | | | | | classe = 2 : 0 (1/0)
| | | | | | classe = 3
| | | | | | | âge < 25,5
| | | | | | | | âge < 23
| | | | | | | | | âge < 15,5 : 0 (1/0)
| | | | | | | | | âge >= 15,5 : 0 (4/2)

| | | | | | | âge >= 23 : 0 (2/0)
```

Le résultat de la classification de l'ensemble de données d'origine est présenté dans le tableau 15. La précision de la classification selon ce tableau est de 77,21 %. Ce résultat est inférieur aux résultats de précision de la classification obtenus dans le tableau 12.

Tableau 15. Matrice de confusion de l'ensemble de données d'origine

Valeur réelle	Valeur prédite		Total réel
	0	1	
0	356	67	423
1	95	193	288
Total	451	260	711
Précision de la classification : 77,21 %			

4. Conclusion

Dans cette étude, l'estimation du nombre de survivants de l'accident du Titanic à l'aide de différentes méthodes a été étudiée. Les facteurs influençant la survie ont été étudiés et le taux de survie a été estimé à l'aide d'une méthode de classification.

Dans un premier temps, des analyses de régression logit et probit ont été réalisées. Ces analyses ont permis de déterminer les variables contribuant de manière significative à la survie et ont révélé une précision de classification de 79,89 % et 79,04 % respectivement. Dans un deuxième temps, deux analyses différentes ont été réalisées à l'aide de l'algorithme d'arbre aléatoire. La première analyse a utilisé les variables utilisées dans les régressions logit et probit qui contribuent de manière significative au modèle. La précision de la classification a été estimée à 81,57 %. La deuxième analyse a été effectuée avec les variables de l'ensemble de données d'origine et la précision de la classification est tombée à 77,21 %.

Lorsque tous les résultats sont pris en compte ensemble, il est préférable d'estimer les données qui contribuent de manière significative au modèle à l'aide d'arbres de décision.

Les résultats de l'étude révèlent qu'en plus des résultats attendus, l'analyse par arbre de décision (exploration de données ou analyse d'apprentissage automatique) avec les données qui contribuent de manière significative au modèle donne des résultats plus fructueux. Ces résultats soulignent que les méthodes d'apprentissage par arbre de décision basées sur les nouvelles technologies sont plus efficaces, mais que les résultats peuvent encore être améliorés par des méthodes statistiques.

Références

[1] E. L. Rasor, « The Titanic: Historiography and Annotated Bibliography ». Greenwood Publishing Group, Londres, 2001.

[2] A. Singh, S. Saraswat, N. Faujdar, « Analyzing Titanic Disaster using Machine Learning ». Conférence internationale sur l'informatique, la communication et l'automatisation, pp. 406-411, 2017.

[3] C. Dieckmann, « The Mystery of the Titanic: What Really Happened ». Undergraduate Research Journal, vol. 13(1), pp. 243-248, 2020.

[4] V. Kshirsagar, N. Phalke, « Analyse de la survie du Titanic à l'aide de la régression logistique ». Revue internationale de recherche en ingénierie et technologie, vol. 6(8), pp. 89-91, 2019.

[5] Kaggle.com, « Titanic Data Set », http://www.kaggle.com/, consulté en octobre 2020.

[6] A. M. Barhoom, A. J. Khalil, B. S. Abu-Nasser, M. M. Musleh, S. S. Abu-Naser, « Prédiction des survivants du Titanic à l'aide d'un réseau neuronal artificiel ». Revue internationale de recherche universitaire en ingénierie, vol. 3(9), pp. 8-12, 2019.

[7] K. Singh, R. Nagpal, R. Sehgal, « Analyse exploratoire des données et apprentissage automatique sur la base de données du naufrage du Titanic ». 10e Conférence internationale sur le cloud computing, la science des données et l'ingénierie, Inde, janvier 2020.

[8] Y. Kakde, Agrawal, S., « Prédiction des survivants du Titanic à l'aide de techniques d'analyse exploratoire des données et d'apprentissage automatique », International Journal of Computer Applications, vol. 179(44), pp. 32-38, 2018.

[9] J. Garrido, J. Zhou, « Crédibilité totale avec les modèles linéaires généralisés et mixtes ». Bulletin ASTIN, vol. 39(1), pp. 61-80, 2009.

[10] T. Koc, M. A. Cengiz, « Genelleştirilmiş Lineer Karma Modellerde Tahmin Yöntemlerinin Uygulamalı Karşılaştırılması ». Karaelmas Science and Engineering Journal, vol. 2(2), pp. 47-52, 2012.

[11] Y. Kida, « Modèles linéaires généralisés : introduction à la modélisation statistique avancée ». Towards Data Science, septembre 2019.

[12] B. Bozkurt, « Kredi ve Yurtlar Kurumunda Kalan Öğrencilerin Memnuniyet Derecelerinin Lojistik Regresyon Yöntemi ile Araştırılması : Edirne Ili Örneği ». Projet de fin d'études du département des sciences sociales de l'université de Trakya, août 2011.

[13] G. Çırak, Ö. Çokluk, « The Usage of Artifical Neural Network and Logistic Regresssion Methods in the Classification of Student Achievement in Higher Education ». Mediterranean Journal of Humanities, vol. 3(2), pp. 71-79, 2013.

[14] D. N. Gujarati, N. C. Porter, « Temel Ekonometri ». Ümit Şenesen ve Gülay Günlük Şenesen (çev.) İkinci Basım, Literatür Yayıncılık, İst. 2001.

[15] Ö. İ. Güneri, B. Durmuş, « Modèles à variables fictives dépendantes : application des modèles logit, probit et tobit aux données d'enquête ». International Journal of Computational and Experimental Science and Engineering, vol. 6(1), pp. 63-74, 2020.

[16] M. Bilki, Ü. Aydın, « Konut Sahibi Olma Kararlarını Etkileyen Faktörler : Lojistik Regresyon ve Destek Vektör Makinelerinin Karşılaştırılması ». Dumlupınar Üniversitesi Sosyal Bilimler Dergisi, vol. 62, pp. 184-199, 2019.

[17] S. Demirci, M. Astar, « Türkiye'de Özel Sigortayı Etkileyen Faktörler: Logit Modeli ». Trakya Üniversitesi Sosyal Bilimler Dergisi, vol. 13 (2), pp. 119-130, déc. 2011.

[18] T. Amemiya, « Qualitative Response Models: A Survey ». Journal of Economic Literature, vol. 19(4), pp. 481-536, 1981.

[19] J. H. Aldric, F. D. Nelson, « Linear Probability, Logit and Probit Models », Sage Publications, États-Unis, 1984.

[20] D. Bertsimas, J. Dunn, « Optimal Classification Trees ». Mach Learn, vol. 106, pp. 1039–1082, 2017.

[21] J. Ali, R. Khan, N. Ahmad, L. Maqsood, « Random Forests and Decision Trees ». International Journal of Computer Science Issues, vol. 9, pp. 5-3, septembre 2012.

[22] G. Nuti, L. A. J. Rugama, « A Bayesian Decision Tree Algorithm ». arXiv:1901.03214v2 [stat.ML], janvier 2019.

[23] B. Gupta, A. Rawat, A. Jain, A. Arora, R. Dhami, « Analysis of Various Decision Tree Algorithms for Classification in Data Mining ». International Journal of Computer Applications, vol. 163 (8), pp. 15-19, avril 2017.

[24] S. D. Jadhav, H. P. Channe, « Étude comparative des techniques de classification K-NN, Naive Bayes et arbre de décision ». International Journal of Science and Research, vol. 5 (1), pp. 1842-1845, janvier 2016.

[25] B. Durmus, Ö. İ. Güneri, « Data Mining with R: An Applied Study ». International Journal of Computing Sciences Research, vol. 3(3), pp. 201-216, 2019.

[26] Ö. Akar, O. Güngör, « Classification d'images multispectrales à l'aide de l'algorithme de forêt aléatoire ». Journal of Geodesy and Geoinformation, vol. 1(2), pp. 139-146, 2012.