Présentation du sujet

2.1. Le parapente et ses risques

2.2. Importance de la gestion des risques dans le parapente

2.3. Base de données des accidents de parapente

Méthodologie

3.1. Exploration des données

Dans un premier temps, nous avons dû nettoyer et préparer la base de données fournie. En effet les saisies manuelles ont amené de nombreuses erreurs de frappes et blancs. Il a donc fallu éliminer toutes les colonnes présentant des données inexactes, incomplètes ou incohérentes.

Dans l’optique de faire un modèle prévisionnel ou du moins de travailler à faire de la prévention, nous avons supprimé toutes les colonnes relatives à des considérations postes accidents.

Il s’agit des colonnes suivantes comme celles relatives à des pratiques différentes ou comme celle qui indique si l’accidenté pense que la météo a eu un impact sur son accident. De telles considérations bien que très intéressantes ne sont pas pertinentes pour décrire la situation qui a causé l’accident.

Après cette phase de nettoyage, nous avons effectué une analyse exploratoire des données (EDA) qui nous a permis de comprendre les caractéristiques des accidents, le sens de chaque information et pour identifier les variables les plus importantes pour la prédiction des accidents. Dans cette démarche les avis des membres de la fédération qui nous accompagnaient ont été très important car cela prend du temps de se plonger dans un domaine étranger.

Deux colonnes de données sortent du lot :

Etat assuré : sur une échelle de 0 à 3 elle dit si les conséquences de l’accident relèvent plus de la bobologie jusqu’au décès.

Cette colonne servira de variable de sortie.

La variable de sortie d'un modèle prédictif est également appelée la variable cible ou la variable dépendante. Elle représente la variable que le modèle tente de prédire en utilisant les informations contenues dans les variables d'entrée ou les variables indépendantes.

3.2. Sélection des variables pertinentes

Nous avons utilisé des techniques statistiques pour sélectionner les variables les plus pertinentes pour la prédiction des accidents de parapente. Nous avons évalué les variables en termes de leur corrélation avec l'accident et leur capacité à prédire l'accident. Les variables les plus pertinentes ont été incluses dans nos modèles prévisionnels.  
Il est vrai que l'utilisation de variables qui sont fortement corrélées avec la variable cible peut améliorer la performance de la prédiction dans certains cas. Cependant, se limiter à la sélection de variables corrélées peut entraîner une sous-représentation des informations contenues dans les données, car il peut y avoir des variables qui ne sont pas corrélées mais qui contiennent des informations importantes pour la prédiction. C’est pourquoi nous avons choisi de garder un large panel de variable d’entrée.

Dans cette phase du travail, nous avons aussi numérisé les données susceptibles de nous intéresser pour la suite en utilisant la légende suivante :

+nouvelles colonnes

Dans le cas des forêts d'arbres aléatoires (la méthode de prédiction choisie), une méthode plus appropriée pour la sélection des variables est la "variable importance". Cette méthode permet de mesurer l'importance relative de chaque variable dans la prédiction de la variable cible en utilisant des métriques telles que le gain d'information et le taux de variation. Les variables qui ont une importance élevée sont conservées pour la prédiction, tandis que celles qui ont une importance faible peuvent être éliminées pour améliorer l'efficacité du modèle.

Nous avons donc écrit un code Python pour utiliser la variable importance en utilisant la bibliothèque Scikit-Learn sur la base de données.

Le résultat est le suivant pour la prédiction de la gravité de l’accident :

AC NOMBRE DE JOUR DEPUIS LE DEBUT DE l’ANNEE 0.083821  
ac altitude 0.082731  
pph altitude 0.081624  
pph age 0.079196  
AC SURFACE AILE 0.076388  
pph distance 0.075189  
PPH NB HEURES PRATIQUE 0.068244  
AC HEURE 0.061662  
AC NIVEAU PROGRESSION VOL LIBRE 0.035424  
AC AILE HOMOLOGATION PARAPENTE 0.032543  
AC TYPE VOL VOL LIBRE 0.031886  
AC 0 BI VOL LIBRE 0.031355  
AC SITE CONNAISSAE VOL LIBRE 0.030759  
AC TYPE AILE VOL LIBRE 0.026032  
AC SELLETTE PROTECTION PARAPENTE 0.025885  
AC EXPERIENCE 0.025370  
Saison 0.024700  
AC ECOLE VOL LIBRE 0.024508  
AC BREVET VOL VOL LIBRE 0.024294  
AC FREQUEE 0.022910  
PPH SEXE 0.011135  
different relief ac pph 0.010203  
AC CONDITION METEO VOL LIBRE 0.010131  
ac plaine 0 montagne 1 0.008934  
pph plaine 0 montagne 1 0.007578  
AC PARACHUTE PRESEE VOL LIBRE 0.005005  
PPH HANDICAPE 0.001522  
AC CASQUE PRESEE VOL LIBRE 0.000970

Les résultats obtenus sont la liste de l'importance relative de chaque variable d'entrée pour le modèle Random Forest. Les valeurs sont comprises entre 0 et 1, et indiquent la contribution de chaque variable à la prédiction de la variable cible : la gravité de l’accident.

Plus la valeur d'importance est élevée, plus la variable est importante pour la prédiction de la variable cible. Dans notre cas, la variable la plus importante est "AC NOMBRE DE JOUR DEPUIS LE DÉBUT DE L'ANNÉE" avec une importance relative de 0.083821, suivie de près par "ac altitude" et "pph altitude" avec des importances relatives de 0.082731 et 0.081624, respectivement.   
Nous ne garderons que les variables avec une importance supérieur à un certain seuil.

Une règle empirique courante est de conserver les variables dont l'importance est supérieure à la moyenne moins deux fois l'écart type des importances. Cette approche permet de conserver les variables qui ont une contribution significative au modèle tout en évitant la surcharge de variables redondantes.

En l’occurrence on a une moyenne m= 0,03571425 et σ= 0,022105. Ainsi pour notre cas nous avons choisi de garder toutes les variables.

3.3. Méthodes d'apprentissage automatique utilisées

3.3.1. Arbres de décision

Les arbres de décision sont une méthode d'apprentissage automatique qui permettent de représenter graphiquement des règles de décision sous forme d'un arbre. Dans notre cas c’est particulièrement pertinent pour fournir un outil de décision visuel et faire de la prévention. Chaque nœud de l'arbre correspond à la donnée variable d'entrée et chaque branche correspond à une valeur de cette variable. Les feuilles de l'arbre correspondent à une prédiction de la variable de sortie : 0,1,2 ou 3.

Dans le cas de la prévention du parapente, les arbres de décision peuvent être utilisés pour prédire le niveau de risque d'un vol en fonction de différentes variables d'entrée telles que les conditions météorologiques, les caractéristiques de l'aile, l'expérience du pilote, etc.

Les données ont été découpées en deux parties : l’ensemble d’entrainement et l’ensemble de teste de portion respective 2/3 et 1/3.

Lorsque la variable de sortie comporte plus de deux classes, on parle de classification multi-classes. Pour construire un arbre de décision pour la classification multi-classes, il existe plusieurs approches possibles :

Approche one-vs-all (ou one-vs-rest) : Cette approche consiste à construire un arbre de décision pour chaque classe en considérant cette classe comme la classe positive et toutes les autres classes comme la classe négative. La prédiction finale est obtenue en sélectionnant la classe pour laquelle l'arbre de décision prédit la probabilité la plus élevée.

Approche multi-étiquettes : Cette approche consiste à considérer toutes les classes simultanément et à construire un arbre de décision qui prédit plusieurs classes en même temps. Dans ce cas, chaque feuille de l'arbre correspond à une combinaison de classes. C’est cette deuxième option que nous avons gardée car elle permet une meilleure lisibilité.

Une fois que l'arbre de décision est construit, il peut être utilisé pour prédire le niveau de risque d'un vol à partir de nouvelles données d'entrée.

Exemple d’arbre de décision :

Une image contenant diagramme

Description générée automatiquement

3.3.2. Forêts d'arbres aléatoires

On utilise la même découpe de la base de données en Train et teste que dans la partie précédente.

Un modèle de forêt d'arbres aléatoires a été construit à l'aide de la bibliothèque scikit-learn. Les hyperparamètres du modèle ont été optimisés en utilisant la validation croisée et la méthode de recherche sur grille (GridSearchCV). Les hyperparamètres optimisés incluent la profondeur maximale des arbres, le nombre minimum d'échantillons requis pour diviser un nœud interne et le nombre minimum d'échantillons requis pour être à un nœud feuille.

Résultats

4.1. Performances des modèles

4.2. Importance des variables

4.3. Identification des circonstances à risque

4.4. Validation des résultats

Discussion

5.1. Interprétation des résultats

5.2. Limitations de l'étude

5.3. Implications pour la gestion des risques en parapente

5.4. Pistes de recherche futures

Conclusion

6.1. Résumé des résultats

6.2. Implications pratiques

6.3. Contribution à la recherche et au développement

Références

7.1. Bibliographie

7.2. Ressources en ligne