Série d'activités d'Actua sur l'IA

Activité 3

L’analyse de régression : faire des prédictions à partir de données

Actua | 111 Murray Street, Ottawa, ON K1N 5M5 | www.actua.ca | 613 234 4137

Jeunesse · STIM · Innovation
Youth · STEM · Innovation

L’analyse de régression : faire des prédictions à partir de données

Vous avez accédé directement à cette activité? Saviez-vous que notre série en compte huit autres? Consultez notre site web pour en connaître les détails et l’ordre recommandé pour les réaliser. Elles présentent toutes des notions d’intelligence artificielle ainsi qu’un vocabulaire propre au domaine. Un glossaire permettant de vérifier le sens des mots nouveaux ou inconnus les accompagne. Amusez-vous tout en réparant une station spatiale!

Toi et ton groupe faites partie d’une équipe d’astronautes et de scientifiques en mission à bord de la station spatiale Actua. Comble de malchance, la station vient d’être bombardée par des rayons magnétiques! Le seul qui peut vous tirer d’affaire, c’est DANN, le réseau neuronal spécialisé d’Actua, mais il a un peu perdu la carte depuis l’incident. Révisez vos connaissances techniques et apprenez quelques notions d’intelligence artificielle pour sauver l’équipage!*

Grâce à vos efforts dans l’activité « Les arbres décisionnels : Classer des objets spatiaux », le classifieur d’objets spatiaux fonctionne de nouveau. DANN procède en ce moment à un diagnostic pour déterminer quelles autres réparations seront nécessaires. Rien ne peut être fait tant qu’il n’aura pas terminé. Entretemps, vous êtes tombés sur les données d’une expérience scientifique menée autrefois à bord de la station dans le but d’étudier l’effet de la microgravité sur les astronautes. Le Centre de contrôle vous a demandé d’y jeter un coup d’œil pour voir s’il est possible d’en faire une analyse préliminaire. Lorsque vous aurez achevé cette tâche, DANN aura terminé son diagnostic et vous pourrez commencer le travail de réparation dans l’activité « Faire le tri : à la découverte de la classification des images »!

**Pour « Dedicated Actua Neural System »*

Présentation de l'activité

Dans cette activité, les élèves analyseront des données pour définir la relation qui unit deux variables et faire des prédictions. Ils feront d'abord une analyse visuelle des données, puis ils appliqueront une méthode appelée « régression » pour créer un modèle capable de faire des prédictions à partir de celles-ci. Enfin, ils réfléchiront à la conception et à l'évaluation des modèles de prédiction et discuteront de l'usage qu'on peut faire de l'apprentissage automatique pour créer ce genre de modèle.

Activité conçue par Actua, en 2022.

Contexte de l'activité	Durée	Public cible
Laboratoire informatique/salle de classe équipée d'ordinateurs	60-75 minutes	Niveaux 9-12 (13-18 ans)

Résultats d'apprentissage

À la suite de cette activité, les participants pourront :

- Analyser la relation entre les variables dépendantes et indépendantes.
- Comprendre l'importance des modèles de prédiction et leur impact.
- Explorer et analyser la précision variable des statistiques et des modèles mathématiques.

OUTILS	COMPÉTENCES	ATTITUDES
Connaissances, ressources et expériences • Analyser la relation entre des variables	**Compétences numériques, compétences en STIM, employabilité et aptitudes essentielles**	**Intelligence numérique, action communautaire et pensée computationnelle**

OUTILS	COMPÉTENCES	ATTITUDES
dépendantes et indépendantes. • Faire des prédictions au moyen de la régression linéaire.	**à la vie** • Analyse statistique • Pensée critique • Conception de diagrammes	• Comprendre l'importance des modèles de prédiction et leur incidence. • Explorer et analyser la précision variable des modèles statistiques et mathématiques.

Logistique (durée, taille du groupe, matériel)

Titre de la section	Durée	Taille du groupe	Matériel
Introduction : Prédire l'avenir	5 min	Toute la classe	
Activité n° 1 : Qu'est-ce que la régression?	15 min	Toute la classe	• Copie de données et du diagramme de dispersion (ou accès par ordinateur) • Stylo/crayon
Activité n° 2 : Créer un modèle	20 min	Chacun pour soi ou en petits groupes	Pour chaque élève ou petit groupe : • Ordinateur avec accès Internet • Diagramme interactif
Activité n° 3 : Valider les prédictions	20 min	Chacun pour soi ou en petits groupes	Pour chaque élève ou petit groupe : • Ordinateur avec accès Internet • Diagramme interactif • Tableau de calcul

Titre de la section	Durée	Taille du groupe	Matériel
			imprimé (ou affiché sur ordinateur) • Stylo/crayon • Calculatrice
Réflexion et récapitulation	10 min	Toute la classe	

Consignes de sécurité

Les consignes de sécurité ci-dessous ne sont pas exhaustives. Veillez à passer en revue l'activité et à inspecter l'environnement où elle sera réalisée afin de déterminer si des mesures additionnelles sont requises pour assurer la sécurité des élèves.

Sécurité en ligne

Certains volets de cette activité nécessitent l'usage d'appareils connectés à Internet.

- Examinez au préalable les vidéos, les sites web et le matériel prévus afin de vous assurer qu'ils conviennent à vos élèves.
- Au besoin, rappelez aux jeunes de se concentrer sur la tâche à faire et d'utiliser uniquement les liens fournis pour l'activité.

Marche à suivre

Introduction : Prédire l'avenir

La prédiction est l'action qui consiste à formuler une hypothèse à propos de l'avenir sur la base de l'information qu'on détient. L'apprentissage automatique peut servir à entraîner des modèles de prédiction pour tout un éventail de tâches ayant un impact énorme sur nos vies. En groupe, réfléchissez aux questions suivantes.

1. Dans quels domaines pourrait-on utiliser des modèles de prédiction dans notre société? Que permettraient-ils de prédire?
    a. Parmi les exemples : les prévisions météorologiques, la propagation des maladies, les éruptions volcaniques, les séismes, les marchés financiers ou boursiers, les migrations animales, le climat.
2. Qu'est-ce qui fait un « bon » modèle de prédiction? Pourquoi est-il important qu'il le soit?
    a. On peut considérer qu'un modèle de prédiction est « bon » si ses prédictions sont justes.
    b. Un modèle qui ferait des prédictions erronées ne serait d'aucune utilité pour la prise de décisions (et pourrait même être dangereux s'il conduisait à de mauvaises décisions).

Activité n° 1 : Qu'est-ce que la régression?

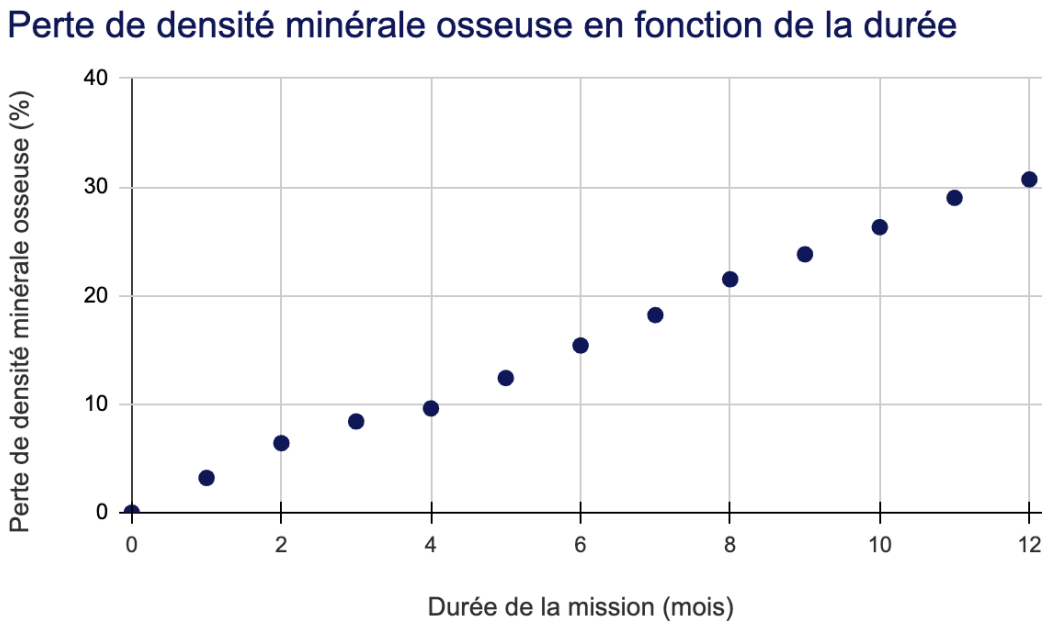
Pour que nos prédictions soient justes, il faut d'abord recueillir des données qui ont un lien avec celles qu'on cherche à étudier. Si on s'intéresse aux effets des longues missions dans l'espace sur la santé humaine, par exemple, il faut recueillir des données sur l'état de santé des astronautes et les analyser dans le but de dégager des tendances. Le tableau ci-dessous présente des données provenant des examens médicaux du dernier groupe d'astronautes à avoir séjourné dans la station spatiale Actua.

Nombre de mois	Perte de densité minérale osseuse (%)
0	0,0
1	3,2
2	6,4
3	8,4
4	9,6
5	12,4
6	15,4
7	18,2
8	21,5

Nombre de mois	Perte de densité minérale osseuse (%)
9	23,8
10	26,3
11	29,0
12	30,7

Ces chiffres sont des moyennes calculées à partir des données recueillies auprès de six astronautes ayant participé à une mission d'un an. Le taux de perte osseuse est relatif à la densité minérale osseuse de départ (c'est-à-dire un taux de densité minérale de 100 % pour chacun des astronautes au mois 0). *Peut-on utiliser ces données pour tenter de prédire ce qui se produirait au cours d'une mission plus longue? Comment?*

Jetons d'abord un coup d'œil aux données :



À partir de ce diagramme de dispersion, on peut supposer qu'il existe une relation entre le temps passé dans l'espace (le nombre de mois en mission) et l'évolution de la santé des astronautes (représenté par la perte de densité minérale). {En petits groupes ou tous ensemble}, réfléchissez aux questions suivantes.

1. Comment peut-on décrire la relation entre la durée de la mission et la perte de densité osseuse?

    a. Les données présentées dans le diagramme de dispersion semblent former une droite. On pourrait donc qualifier cette relation de linéaire.

2. Pourrait-on utiliser ces données et le diagramme pour prédire l'évolution de la santé des astronautes au bout de 12 mois? Comment?

    a. On pourrait tracer une droite parmi les points de données et extrapoler les résultats à partir du 12e mois.

3. À votre avis, quelle serait la perte de densité minérale au bout de 18 mois? Au bout de 24 mois?

    a. Si l'hypothèse d'une relation linéaire est exacte, on peut prédire le pourcentage de perte en calculant la différence entre les 6e et 12e mois (15 % environ) et en ajoutant ce chiffre à celui du 12e mois (30 + 15 = 45) pour connaître le pourcentage de perte au bout de 18 mois. On procéderait de la même façon pour le calculer au bout de 24 mois (45 + 15 = 60).

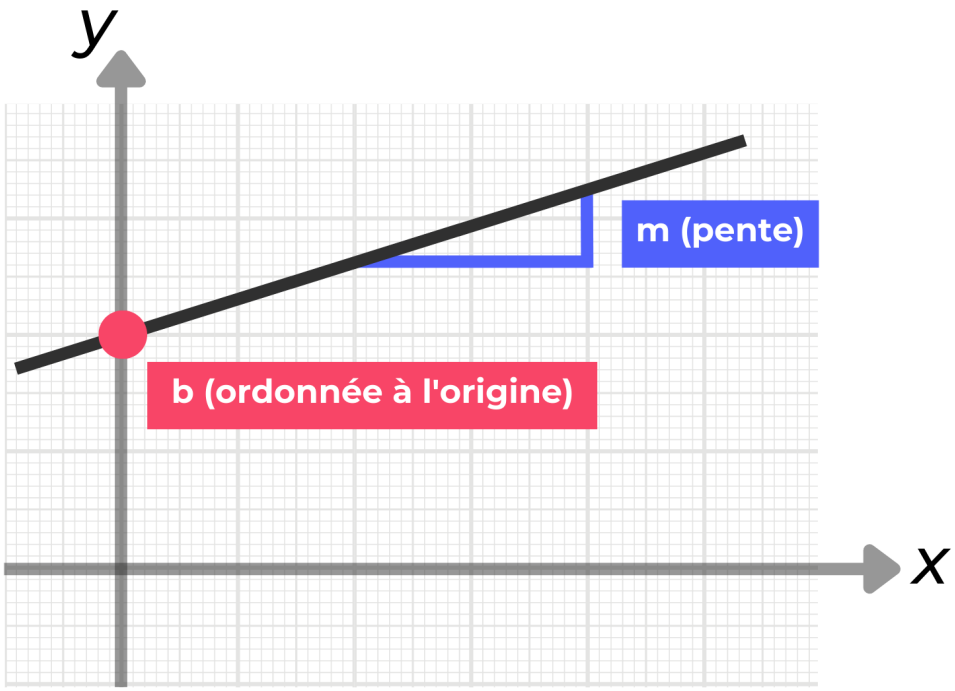
    b. Pour prédire le pourcentage de perte au bout de 24 mois, on peut aussi calculer la différence entre le mois 0 et le 12e mois et ajouter une valeur identique (30 + 30 = 60).

Activité n° 2 : Créer un modèle

Notre diagramme de dispersion nous a permis de réaliser des estimations « à vue de nez », mais si on voulait exprimer mathématiquement la relation entre ces deux variables, comment faire? On peut recourir à une méthode appelée « régression ». La régression permet de créer un modèle qui décrira la relation entre le temps passé dans l'espace et son effet sur la perte de densité osseuse. Puisqu'il s'agit d'une relation linéaire, nous ferons appel à la régression dite « linéaire ». Il existe aussi d'autres types de régression qui servent à modéliser des données plus complexes. Voici comment il faut procéder :

1. Tracez une droite en tentant de la faire passer au plus près de tous les points de données.
2. Calculez le degré de « précision » du tracé en comparant la distance entre chaque point et la droite.
3. Ajustez le tracé de la droite et répétez l'étape 2 afin de vérifier si vous avez obtenu ou non un meilleur alignement par rapport aux données.
4. Répétez l'étape 3 jusqu'à ce que vous ne soyez plus capables d'améliorer l'alignement de la droite.

Rappelez-vous que l'équation d'une droite représentant une relation linéaire s'écrit comme suit : **y = mx + b**



$y = mx + b = 0{,}5x + 2$

où **m** correspond à la pente de la droite et **b** à la hauteur de la droite (l'« ordonnée à l'origine », soit la valeur de y lorsque x = 0). Dans le cas d'une droite de régression, **y** devient **ŷ** pour indiquer qu'il s'agit d'une valeur prédite, plutôt que d'une valeur établie. Dans notre modèle, on pourrait l'écrire de la manière suivante :

(ŷ : perte de densité osseuse prédite) = (pente : taux de perte) × (x : mois dans l'espace) + b

Les deux termes soulignés, *pente* et *b*, correspondent aux valeurs qu'on peut modifier pour ajuster la droite de régression. La pente a une incidence sur « l'angle » de notre droite et **b** la fait monter et descendre dans le diagramme. « Mois dans l'espace » est notre variable indépendante; en lui attribuant une valeur, on peut se servir de l'équation afin d'obtenir une valeur pour la « perte de densité osseuse prédite », soit notre variable dépendante.

Nous sommes maintenant prêts à choisir des valeurs de départ pour la pente et pour b. On pourrait choisir n'importe lesquelles, mais ce serait probablement une bonne idée de commencer avec les données du tableau. Si on trace une droite traversant le premier point de données (0,0) et le dernier (12, 30,7), on calculera la pente de la façon suivante :

(y2 - y1) / (x2 - x1) = (30,7 - 0) / (12 - 0) = 2,56

De la même façon, si on utilise le premier point de données (0, 0), on établira la valeur de **b = 0**. Notre premier modèle ressemblera donc à ceci :

(ŷ : perte de densité osseuse prévue) = 2,56 × (x : mois dans l'espace) + 0

Vous pouvez toutefois utiliser n'importe quels deux points de données afin de calculer vos propres valeurs pour la pente et pour b. Dans le diagramme interactif ci-dessous, placez deux points en cliquant à deux endroits à l'intérieur de la grille. Vous obtiendrez ainsi la valeur de la pente et la valeur de b pour la droite ainsi tracée. Vous remarquerez aussi la présence de traits en jaune : ils représentent les « résidus » pour chacun des points de donnée. La question des résidus sera abordée dans la prochaine partie.

Diagramme interactif

Activité n° 3 : Valider des prédictions

Nous connaissons maintenant l'équation d'un modèle linéaire : que faut-il faire ensuite? Un modèle prédictif ne sert pas à grand-chose si ses prédictions ne correspondent pas à la réalité. Il existe quelques méthodes pour en vérifier la justesse; celle que nous utiliserons dans cette activité pour comparer différentes droites du modèle s'appelle la « somme des carrés des résidus (SCR) ».

Qu'est-ce qu'un « résidu »?

Dans la méthode de la régression linéaire, un résidu correspond à l'écart entre une valeur observée, comme celles de notre jeu de données, et la valeur prédite par notre modèle. La valeur du résidu nous indique dans quelle mesure notre prédiction est juste ou erronée. Plus notre prédiction se rapproche de la valeur observée dans nos données, et plus la valeur du résidu est faible. On la calcule de la façon suivante :

(résidu) = (valeur observée) - (valeur prédite)

La méthode appelée *somme des carrés des résidus* additionne les carrés des résidus (c.-à-d. résidu × résidu) calculés pour chaque point de données. Plus la somme des carrés des résidus est faible, et plus les prédictions du modèle sont proches des valeurs observées. Entre deux modèles, celui qui obtient la somme des carrés la plus faible est donc le meilleur pour faire des prédictions.

En utilisant le modèle présenté plus haut,

($\hat{y}$: perte de densité osseuse prévue) = 2,56 × (x : mois dans l'espace) + 0

calculez les carrés des résidus pour les points de données restants, puis calculez la somme des carrés de résidus.

Mois en mission	Perte de densité osseuse	Perte de densité osseuse prédite par le modèle	Résidus (observés - prédits)	Carrés des résidus
0	0,0	0,0	0,0	0,0
1	3,2	2,6	0,6	0,4
2	6,4	5,1	1,3	1,7
3	8,4	7,7	0,7	0,5
4	9,6	10,2	-0,6	0,4
5	12,4	12,8	-0,4	0,2
6	15,4	15,4	0,0	0,0
7	18,2	17,9	0,3	0,1

8	21,5	20,5	1,0	1,0
9	23,8	23,0	0,8	0,6
10	26,3			
11	29,0			
12	30,7			

Pour calculer la somme des carrés des résidus, on additionne toutes les valeurs dans la colonne de droite. Malheureusement, ce chiffre n'est pas très utile en soi. Il indique plutôt qu'il faut refaire les mêmes calculs encore et encore jusqu'à ce qu'on arrive au bon modèle (c'est-à-dire lorsque les valeurs de la pente et de b donneront la somme des carrés des résidus la plus faible possible).

Diagramme interactif

Revenons au diagramme interactif. Vous verrez affichée dans le rapport la somme des carrés des résidus pour une droite donnée (après les lettres SSR, pour « sum of squared residuals »). Les traits jaunes reliant les points de données à la droite d'ajustement en vert correspondent aux résidus : le but est d'arriver à ce que la somme de leurs carrés donne la plus petite valeur possible.

1. Explorez le diagramme interactif durant quelques minutes.
2. Quelle est la plus faible valeur de SCR que vous pouvez trouver? Quelle est l'équation de ce modèle?
3. Choisissez le meilleur modèle pour prédire le taux de perte de densité osseuse au bout de 18 mois et de 24 mois. Les prédictions se comparent-elles à vos estimations de départ?

Réflexion et récapitulation

Tout comme les arbres décisionnels, la régression offre un exemple d'application possible de l'apprentissage automatique. {En petits groupes ou tous ensemble}, réfléchissez aux tâches que vous avez accomplies dans cette activité et discutez des questions suivantes.

1. De quelle manière peut-on associer l'apprentissage automatique et l'analyse de régression?

    a. Les modèles de prédiction fondés sur une analyse de régression comportent des paramètres qu'il est possible d'ajuster pour en améliorer la précision (comme la pente et l'ordonnée à l'origine, mais d'autres aussi, dans les modèles non linéaires).

    b. Si l'on se fixe un but comme trouver la plus petite somme des carrés des résidus, l'apprentissage automatique peut nous aider à trouver les meilleurs paramètres pour un jeu de données particulier et à refaire nos calculs lorsque de nouvelles données s'y rajoutent.

2. Pourquoi les modèles de régression conviennent-ils particulièrement bien à l'apprentissage automatique?

    a. Il faut faire un grand nombre de calculs pour trouver les meilleures valeurs pour les paramètres d'un modèle. Les ordinateurs sont beaucoup plus rapides sur ce plan que les êtres humains. Ce facteur n'est pas à négliger lorsque la quantité de données ou le nombre de variables augmentent.

3. Quels facteurs ont un effet sur la justesse des prédictions de notre modèle?

    a. Puisque nous avons établi les paramètres du modèle à partir des données dont nous disposions, nous avons supposé qu'elles étaient représentatives, c'est-à-dire que les données observées étaient caractéristiques de l'astronaute moyen. Si les données n'étaient pas représentatives, le modèle ne donnerait pas de bons résultats à l'extérieur de ce groupe d'astronautes.

    b. Même si les données recueillies semblent indiquer une relation linéaire pour la période de 0 à 12 mois, on ne peut pas nécessairement conclure que ce sera vrai pour les périodes subséquentes (18 et 24 mois). La relation effective entre la variable indépendante (nombre de mois dans l'espace) et la variable dépendante (perte de densité osseuse) pourrait s'exprimer sous une autre forme, comme une courbe polynomiale ou logistique.

4. Comment faire pour améliorer les prédictions d'un modèle?

    a. Une plus grande quantité de données permettrait d'améliorer les prédictions du modèle, car nous serions davantage certains que les données qui servent à entraîner ses paramètres sont représentatives de l'effet réel de la variable indépendante, plutôt que de s'appliquer uniquement aux astronautes de notre échantillon.

Même si les données fournies dans le tableau ne sont pas de vraies données expérimentales, elles se fondent sur des expériences bien réelles menées sur des astronautes. Saviez-vous que leur masse osseuse diminue de 1 % à 2 % par mois sous l'effet de la microgravité (c.-à-d. dans l'espace)? Pour en apprendre davantage, visitez le site de l'Agence spatiale canadienne : https://www.asc-csa.gc.ca/fra/sciences/mso/os.asp.

Possibilités d'adaptation

Il est possible d'adapter différents aspects de cette activité (durée, environnement, matériel, taille du groupe ou instructions) pour la rendre plus accessible ou plus complexe. Les **modifications** ci-dessous vous permettront de diminuer le niveau de difficulté de l'activité et les **ajouts**, d'augmenter sa durée ou son niveau de difficulté.

Ajouts

- Demandez aux élèves de choisir ou de créer leur propre jeu de données à deux variables, par exemple, la longueur de jambe par rapport à la vitesse de course, ou la taille par rapport à l'envergure des bras. Dites-leur d'insérer leurs données dans le diagramme interactif et de trouver le meilleur modèle possible (c.-à-d. la plus faible valeur de SCR possible).
- Demandez aux élèves de formuler des questions au sujet des données présentées dans l'activité ou de leurs propres données. Voici un exemple : « Quelle serait la perte de densité osseuse au bout de 13,5 mois? » ou « Quelle serait l'envergure des bras d'une personne de 7 pieds? ».

Modifications

- Pour réduire la complexité de l'activité, sautez les parties qui exigent des calculs à la main et concentrez-vous sur les modèles du diagramme interactif.
- Pour pousser plus loin l'analyse des données interactives, demandez aux élèves de créer leurs propres diagrammes dans Google Sheets. Comment les données sont-elles représentées selon différents diagrammes? Quels diagrammes pourraient-ils faciliter l'analyse par les modèles de prédiction?

Références et remerciements

Plusieurs exercices ont été conçus à l'aide de p5.js, une bibliothèque JavaScript en ligne accessible à : https://p5js.org/.

Agence spatiale du Canada (ASC, 2006). *Comment les os réagissent-ils dans l'espace?* Source : https://www.asc-csa.gc.ca/fra/sciences/mso/os.asp.

Conditions d'utilisation

Avant de réaliser cette activité en tout ou en partie, vous reconnaissez et acceptez ce qui suit :

- Il vous appartient de passer en revue toutes les sections du présent document et la documentation connexe ainsi que d'appliquer les consignes de sécurité nécessaires à la protection de toutes les personnes concernées;
- Les mesures précisées à la rubrique « Consignes de sécurité » du présent document ne sont pas exhaustives ni ne remplacent votre propre cadre d'examen de la sécurité;
- Actua n'est pas responsable des dommages attribuables à l'usage du présent contenu;
- Vous pouvez adapter ce document à vos besoins (le remanier, le transformer ou créer du matériel à partir de celui-ci), à condition d'indiquer qu'Actua en est l'auteur original et que vous y avez apporté des changements. Ce contenu ne peut être transmis à de tierces parties sans la permission écrite d'Actua.

À propos d'Actua

Représentant plus de 40 universités et collèges à travers le pays, Actua est le principal réseau de sensibilisation des jeunes aux sciences, à la technologie, à l'ingénierie et aux mathématiques (STIM) au Canada. Chaque année, 350 000 jeunes prennent part à des ateliers pratiques, à des camps et à des projets communautaires inspirants dans plus de 500 localités d'un océan à l'autre. Actua met l'accent sur la participation de jeunes sous-représentés dans le cadre de programmes destinés aux Autochtones, aux filles et aux jeunes femmes, aux jeunes à risque ainsi qu'à ceux vivant dans des communautés nordiques ou éloignées. Pour de plus amples renseignements, consultez notre site web à actua.ca et suivez-nous sur Twitter, Facebook, Instagram et YouTube!

Annexes

Annexe A – Liens carrières/mentorat :

- Chercheur/chercheuse en apprentissage automatique
- Programmeur/programmeuse
- Ingénieur/ingénieure logiciel
- Chercheur/chercheuse