Cahier des charges
du projet de formation

en référence à la norme AFNOR NF X 50 769

DATA SCIENCE :
Machine learning et Data/text mining

Présentation de l'ANSES

L'Agence Nationale chargée de la sécurité sanitaire de l'alimentation, de l'environnement et du travail *assure des missions de veille, d'expertise, de recherche et de référence sur un large champ couvrant la santé humaine, la santé et le bien-être animal et la santé végétale. Elle offre une lecture transversale des questions sanitaires.*

L'Agence couvre ainsi de manière globale l'ensemble des expositions (particules, ondes, inhalation, ingestion...) auxquelles un individu peut être sujet, volontairement ou non, à tous les âges et moments de sa vie qu'il s'agisse d'expositions au travail, pendant ses transports, ses loisirs, ou via son alimentation.

L'Anses évalue également de manière transverse les risques et les bénéfices sanitaires en y intégrant l'apport des sciences humaines et sociales, transmet ses avis et recommandations aux pouvoirs publics et rend systématiquement publics ses travaux.

C'est un établissement public à caractère administratif placé sous la tutelle des ministères chargés de la Santé, de l'Agriculture, de l'Environnement, du Travail et de la Consommation.

L'Agence compte près de 1350 agents et mobilise environ 800 experts extérieurs via ses collectifs d'experts.
Elle s'appuie sur un réseau de 11 laboratoires de référence et de recherche, répartis sur 16 implantations géographiques sur le territoire et reconnus au niveau international dans plusieurs domaines ou disciplines.

SOMMAIRE

I. Contexte

A. Situation à l'origine de la demande

La Direction de l'Evaluation des Risques (DER) du pôle « Sciences pour l'expertise » assure l'ensemble des missions d'évaluation dans le domaine des bénéfices/risques nutritionnels et sanitaires liés à l'alimentation, des risques liés à la santé-environnement, des risques liés à la santé au travail, liés à la santé, à l'alimentation et au bien-être des animaux, à la santé des végétaux.
Elle fait appel des comités d'experts spécialisés (CES) et autres collectifs d'experts constitués auprès de l'Agence en coordonnant leurs travaux.
Elle fait également appel aux compétences scientifiques de ses personnels et travaille en liaison avec les autres entités de l'Agence.
Dans son domaine de compétences, elle assure certaines missions d'alerte et de vigilance, organise les études et enquêtes nécessaires à la collecte de données utiles à ses travaux d'expertise. Elle gère les observatoires et bases de données qui y sont associées et mène à bien les développements méthodologiques nécessaires.

Ces derniers temps, s'opère une révolution dans la manière d'analyser, base par base ou conjointement, des données complexes et de sources différentes et les équipes de l'ANSES se doivent de développer leurs compétences dans ce domaine par la veille sur les différents outils, et développer leur capacité d'agir.

B. Résultats attendus

Les résultats attendus pour cette formation sont :

- Gagner en efficience : gain de temps, et optimisation des analyses de données pour en tirer les résultats et renseignements les plus pertinents
- Accroître les potentialités par la connaissance des méthodes et outils les plus pertinents du domaine : identifier les possibilités et limites des méthodes et outils
- Internaliser l'expertise dans ce domaine d'activité : développer les compétences d'une communauté, favoriser l'entraide

II. Le public

Cette formation s'adresse à une communauté intra-direction, jusqu'à 10 scientifiques (au moins Bac + 5) de l'Anses.
Les personnes visées ont de bonnes bases en statistique inférentielle, en particulier en analyse discriminante par arbres de décision ou régression logistique et en régression linéaire, ainsi que de bonnes notions de programmation avec R.

III. Particularités de la formation attendue

A. Objectifs de la formation

Comprendre ce qu'est le Machine Learning (définitions, positionnement par rapport aux statistiques inférentielles, lien avec data mining), la nature des problèmes qu'il permet de résoudre. Apprendre à mettre en œuvre les principaux algorithmes sur des données réelles, à analyser leurs résultats, et comparer leurs performances.

Mettre en œuvre les méthodes de la statistique textuelle sur des corpus de nature différente (questions ouvertes, entretiens, mots associés, articles de presse, pages Web, etc.) à l'aide de logiciels spécifiques. Interpréter et présenter les résultats.

IMPORTANT : Les apprentissages et mises en œuvre des principaux algorithmes doivent être traités et présentés à travers des études de cas, si possible dans le domaine de la santé.

B. Axes de programme

Machine learning (bagging, boosting, SVM, optimisation...) : a priori 3 jours

Définitions, positionnement par rapport aux statistiques inférentielles, lien avec data mining, et les problématiques qu'il permet de résoudre. Apprentissage et mise en œuvre des principaux algorithmes, analyse, interprétation et comparaison des résultats, performances, comparaison de modèles, limites. Introduction au deep learning et à ses algorithmes et applications.

a. Méthodes d'agrégation ou méthodes d'ensemble
b. Machine learning : Support Vector Machines et méthodes à noyaux
c. Deep learning et ses applications (introduction) : les réseaux de neurones

Contenu :

Méthodes d'agrégation ou méthodes d'ensemble

- Bagging et forêts aléatoires : réduction de variance;
- Mesures d'importance des variables
- Agrégation d'arbres de décision
- Boosting : réduction de biais

Machine learning : Support Vector Machines (SVM) et méthodes à noyaux

- Support Vector Machines linéaires
- Support Vector Machines non linéaires
- Astuce du noyau et généralisation

Deep learning et ses applications (introduction) : réseaux de neurones

- Comprendre le transfer learning
- Les principaux environnements : Tensorflow, Keras
- Les différents types de réseaux : Réseaux convolutionnels Réseaux récurrents

Text Mining avec des statistiques textuelles : a priori 2 jours

Méthodes de la statistique textuelle (text mining) sur des corpus de textes (questions ouvertes, entretiens, mots associés, articles de presse) et leur mise en œuvre : l'analyse, l'interprétation et la présentation des résultats.

a. Les principes du traitement du langage naturel, de l'analyse textuelle et du text mining
b. Préparation des données selon le problème abordé
c. Traitement d'un corpus textuel par les méthodes de statistique textuelle

Contenu :

Les principes de l'analyse textuelle et le text mining

- Qu'est-ce que le traitement du langage naturel ?
- A quelles questions répond le traitement du langage naturel ?
- Quels apports de la statistique textuelle (text mining) par rapport à l'analyse qualitative
- Les logiciels d'aide à la lecture de textes.

Préparation des données textuelles en fonction de la problématique

- Données d'enquêtes
- Automatisation de processus

Traitement d'un corpus par les méthodes de statistique textuelle

- Construction du lexique et préparation des tableaux
- Lemmatisation ou pas ?
- Application des algorithmes de machine learning
- Les possibilités pour le traitement du langage avec le deep learning
- Production des statistiques uni ou multivariées associés.

Interpréter les résultats

- Mots du lexique, concordances, mots spécifiques, plans factoriels et arbres de classification

C. Durée

A proposer par le prestataire, environ 5 jours total. La formation sera réalisée en deux modules distinctes selon les deux axes présentés ci-dessus, espacés d'un à deux mois.

D. Calendrier

Second semestre 2020

E. Modalités pédagogiques

Les stagiaires seront encouragés à s'entraîner entre les séances. Pour cela, il leur sera proposé un support pédagogique adapté et des activités complémentaires possibles (revues, vidéos, etc …) leur seront conseillées.
L'ANSES souhaite disposer de supports de formation qui seront remis au cours de la formation à chaque stagiaire.

Technologies : les logiciels ou langages utilisés lors de la formation doivent être en accès libre (gratuit), R étant le logiciel statistique de prédilection utilisé à l'Anses.

F. Modalités d'évaluation

1. Audit de positionnement

Le prestataire retenu pourra proposer un questionnaire de positionnement afin de s'assurer par exemple des pré-requis nécessaires à la formation, ou recueillir les attentes des apprenants.

2. Evaluation des acquis

Une évaluation pourra être proposée par le prestataire, en cours ou en fin de formation.

3. Evaluation des compétences

L'évaluation des compétences aura lieu dans l'année suivant la formation, lors de l'entretien annuel.

4. Evaluation de la satisfaction

La satisfaction sera évaluée par un questionnaire à l'issue de la formation par le prestataire, ainsi que par l'ANSES.

G. Modalités d'organisation

1. Logistique

- La formation se déroulera en présentiel avec un formateur, et seront organisées dans les locaux de l'Anses (site de Maisons-Alfort), chaque apprenant devant être muni d'un ordinateur portable pour la réalisation des exercices pratiques.

Le prestataire s'engage à pouvoir dispenser en présentiel des séances de formation au sein des entités de l'Agence, et de s'assurer que les moyens, matériels notamment, nécessaires à la formation soient disponibles à cet effet. Si tel n'est pas le cas, le prestataire s'engage à mettre à disposition le matériel nécessaire.

Les horaires habituels de formation sont 9h30-17h30, avec une heure de pause déjeuner.

2. *Volume prévisionnel*

Une session de 10 apprenants maximum.

3. *Calendrier prévisionnel*

Le démarrage de la prestation débutera dès que possible.

4. *Les supports*

Un exemplaire de l'ensemble des supports pédagogiques et des supports d'évaluation proposés aux stagiaires devra être fourni pour validation à la DRH de l'ANSES.

Pour les apprenants, les supports pourront être mis à disposition par voie électronique ou en version papier, envoyés sur le site de déroulement de la formation au plus tard 8 jours avant la formation.
La mise à disposition et la reproduction des supports, s'il y a lieu, sont à la charge du prestataire.

H. **Modalités de suivi**

1. *Suivi pédagogique et évaluation du dispositif*

Une réunion de cadrage sera réalisée avec le prestataire et les commanditaires du projet en amont du démarrage de la prestation, afin de :
- Valider les propositions du prestataire et adapter si nécessaire,
- Préciser l'organisation de la prestation.

2. *Suivi administratif et financier*

a) *Budget prévisionnel*

L'achat se fait dans le cadre d'une mise en concurrence simple, avec un budget inférieur à 40 000 euros HT selon le code des marchés publics.
Le budget prévisionnel pour ce projet est d'environ 2000 euros HT par jour de formation.

b) *Les bons de commande*

Les commandes seront réalisées dés validation de la prestation par le service SDRH.

c) *Les attestations de formation/de présence*

Les attestations de présence sont établies par le prestataire.

Les attestations de formation sont établies par le prestataire en fin de parcours.

d) *La facturation*

La facture devra mentionner :
- L'intitulé de la formation suivie et les dates
- La référence ANSES de la commande

- Pour la facturation, les documents dématérialisés (facture et attestations) devront être :
- Adressés par mail au référent formation de l'ANSES, ou déposé sur un espace dédié ;
- Mis à disposition sous CHORUS, avec une organisation par dossier par facture comprenant la facture et les attestations de présence correspondantes.

- La mise en paiement des factures sera réalisée après constatation du service fait sur présentation des attestations de présence.

7. Adaptation de la réponse

Une adaptation du contenu des formations pourra être demandée, sans surcoût, en réponse à un besoin spécifique à un métier ou un sujet professionnel.
Les modalités pédagogiques proposées resteront dans le cadre des modalités identifiées dans ce document.