Université du Québec en Outaouais **Département d'informatique et d'ingénierie**

Sigle : INF5173 Gr. 01
Titre : Fouille et entreposage de données
Session : Automne 2024 Horaire et local
Professeur : Tajeuna, Etienne

1. Description du cours paraissant à l'annuaire :

Objectifs

Permettre aux étudiantes et étudiants de maîtriser les fondements, les concepts et les problèmes liés à la fouille et à l'entreposage de données à des fins de prise de décision. Les aspects de la visualisation de l'information et de la connaissance seront également présentés tel que requis en intelligence d'affaires (veille économique ou Business intelligence).

Contenu

Fouille de données : étapes de découverte de connaissances (prétraitement, fouille de données et interprétation des résultats), techniques de classification (arbres de décision, réseau de neurones, etc.), techniques de regroupement (treillis de concepts, classification hiérarchique), règles d'association, motifs séquentiels, cas aberrants et fouille de données complexes (Web, texte et graphe). Entreposage de données : étapes de construction d'un entrepôt de données (acquisition, stockage, traitement et accès), modélisation multidimensionnelle des données, création de cubes de données, techniques OLAP, types d'architectures des entrepôts de données, optimisation des performances et matérialisation de cubes de données. Visualisation de l'information et de la connaissance. Intégration des deux technologies de fouille et d'entreposage de données. Applications et outils comme ETC (extraction, transformation et chargement).

Descriptif – Annuaire

2. Objectifs spécifiques du cours :

- Familiariser l'étudiant(e) avec les concepts et techniques de la prospection et de l'entreposage des données.
- Présenter les principales techniques de collecte, de préparation, d'exploration, de modélisation et d'entreposage des données.
- Illustrer ces techniques à l'aide de logiciels et outils tels que Python et SQL pour le prétraitement et la prospection des données, ainsi que des outils de Business Intelligence pour l'entreposage et l'analyse des données.
- Permettre à l'étudiant(e) de mettre en pratique toutes les connaissances et techniques acquises durant le cours via des exercices pratiques et des projets durant la session portant sur l'ensemble du processus de prospection des données, de l'acquisition à l'analyse finale.

3. Stratégies pédagogiques :

Les formules pédagogiques suivantes seront utilisées :

Logistique du cours

- Accès à Moodle sur le Web pour la récupération des notes de cours, des énoncés de travaux, des consignes spécifiques, des soumissions des travaux et des résultats d'évaluation.

Plan synthétisé du cours

Les thèmes suivants seront étudiés :

- Techniques de prétraitement de données : *1- Épuration. 2 - Intégration et transformation. 3 - Sélection et réduction*
- Entreposage de données : raison d'être et concepts
- Étapes de construction d'un entrepôt de données

- Modélisation multidimensionnelle
- Stratégies de conception
- Étapes du processus de découverte de connaissances
- Survol des techniques de prospection de données et des applications courantes
- Classification, régression et prédiction : *1- Définitions, principaux thèmes. 2 - Comparaison entre la classification, la régression et la prédiction. 3 - Arbres de décision et règles de classification. 4 - Réseaux bayésiens. 5 - Approches statistiques de prédiction (modèles de régression)*
- Regroupement : méthodes hiérarchiques comme *K-Means*
- Règles d'association et mesures de qualité
- Forage de texte
- Forage web
- Analyse des réseaux sociaux

4. Heures de disponibilité ou modalités pour rendez-vous :

Avant et après le cours ou sur rendez-vous.

Courriel : etiennegael.tajeuna@uqo.ca

5. Plan détaillé du cours sur 15 semaines :

Semaine	Thèmes	Dates
1	**Chap. 1- Introduction à la prospection des données** • Définition et importance de la prospection des données • Applications et cas d'utilisation • Cycle de vie de la prospection des données	03 sept. 2024
2	**Chap. 2- Préparation et exploration des données** • Collecte des données • Nettoyage des données • Transformation et normalisation des données • Gestion des données manquantes et des outliers • Techniques de visualisation des données • Analyse descriptive • Détection des tendances et des patterns • Techniques de réduction de dimensionnalité (PCA, t-SNE, UMAP) • Mesures de similarités • Introduction de RapidMiner • Exercices	10 sept. 2024
3	**Chap. 2- Préparation et exploration des données (suite et fin)** **Présentation projet P1.**	17 sept. 2024
4	**Chap. 3- Entreposage des données** • Introduction à l'entreposage des données • Conception de l'entrepôt de données • Architecture d'un entrepôt de données • ETL (Extraction, Transformation, Chargement) • Modélisation dimensionnelle (schémas en étoile et en flocon)	24 sept. 2024
5	**Chap. 3- Entreposage des données (suite et fin)** **Chap. 4- Analyse des associations** • Règles d'association • Algorithmes d'extraction des règles (Apriori, FP-Growth) • Exercices	01 oct. 2024

	Présentation projet P2; Remise projet P1.	
6	**Préparation examen de mi-session**	08 oct. 2024
7	**Semaine d'étude**	15 oct. 2024
8	**Examen de mi-session**	22 oct. 2024
9	**Chap. 4- Analyse des associations (suite et fin)** **Chap. 5- Techniques de regroupement (Clustering)** • K-means • Hiérarchique • DBSCAN • Évaluation et validation des clusters • Regroupement des données de grandes dimensions • Regroupement basé sur les structures de graphe (Graph clustering) • Exercices **Remise projet P2.**	29 oct. 2024
10	**Chap. 5- Techniques de regroupement (Clustering) (suite et fin)** **Chap. 6- Techniques de classification et régression** • Algorithmes de classification (k-NN, SVM, Naïve Bayes, etc.) • Évaluation et validation des modèles de classification • Méthodes de sélection de modèles • Régression linéaire et non-linéaire • Régression logistique • Évaluation et validation des modèles de régression • Exercices	05 nov. 2024
11	**Chap. 6- Techniques de classification et régression** **Chap. 7- Techniques avancées de prospection des données** • Forage de texte • Analyse des réseaux sociaux • Forage web • Exercices	12 nov. 2024
12	**Chap. 8- Techniques avancées de prospection des données (suite)** **Présentation projet P3.**	19 nov. 2024
13	**Chap. 8- Techniques avancées de prospection des données (suite et fin)**	26 nov. 2024
14	**Préparation examen final**	03 déc. 2024
15	**Examen final** **Remise projet P3.**	10 déc. 2024

6. Évaluation du cours :

L'étudiant(e) dans ce cours sera évalué(e) par les examens de mi-session et final, ainsi que par des travaux pratiques. La pondération de la note finale sera comme suit :

- Examen de mi-session : **25 %**

- Examen final : **35 %**
- Projets (03): **40 % (10 % + 10 % + 20 %)**

Les travaux pratiques comprendront les volets suivants : entreposage de données (01 TP), prétraitement des données (01 TP), regroupement de données (01 TP) et la prédiction (01 TP).

Une moyenne générale inférieure à **64 %** est éliminatoire et conduit automatiquement à l'échec de l'étudiant(e). Les projets se feront par des équipes de deux à trois étudiants et le libellé du projet **PX**(= 1, 2 ou 3) par l'équipe **N** doit être **INF5173-PX-EquipeN**. La pénalité de retard pour la remise d'un travail est de **2 points** par jour (y compris les jours fériés et les fins de semaine).

Des consignes sur l'échéancier et la réalisation des projets seront précisées.

Des consultations de groupes seront organisées sur rendez-vous afin de guider et d'orienter les étudiant(e)s dans la réalisation de leurs travaux.

7. Politiques départementales et institutionnelles :

- Politique du département d'informatique et d'ingénierie relative à la tenue des examens
- Note sur le plagiat et sur la fraude
- Politique relative à la qualité de l'expression française écrite chez les étudiants et les étudiantes de premier cycle à l'UQO
- Absence aux examens : cadre de gestion, demande de reprise d'examen (formulaire)

Tolérance ZÉRO en matière de violence à caractère sexuel.

Le Bureau d'intervention et de prévention en matière de harcèlement (BIPH) a pour mission d'accueillir, soutenir et guider toute personne vivant une situation de harcèlement, de discrimination ou de violence à caractère sexuel. Le BIPH oriente ses actions afin de prévenir les violences à caractère sexuel pour que nous puissions étudier, travailler et s'épanouir dans un milieu sain et sécuritaire.

Vous vivez ou êtes une personne témoin d'une situation de violence à caractère sexuel ? Vous êtes une personne membre de la communauté étudiante ou une personne membre du personnel, autant à Gatineau qu'à Ripon et St-Jérôme, l'équipe du BIPH est là pour vous, sans jugement et en toute confidentialité.

Ensemble, participons à une culture de respect.

Pour de plus amples renseignements consultez UQO.ca/biph ou écrivez-nous au Biph@uqo.ca

8. Principales références :

Notes de cours disponibles sur Moodle (principale référence)

1. Bertrand Burquier. *Business intelligence avec SQL Server 2008*. Mise en œuvre d'un projet décisionnel, Dunod, 2009.
2. Sébastien Fantini. Business Intelligence avec SQL Server 2019 et Power BI - Maîtrisez les concepts et réalisez un système décisionnel, ENI, Mars 2020.
3. Matteo Golfarelli & Stefano Rizzi. Data Warehouse Design: Modern Principles and Methodologies, McGraw-Hill, 2009.
4. Kellyn Gorman, Allan Hirt, Dave Noderer, Mitchell Pearson, James Rowland-Jones, Dustin Ryan, Arun Sirpal, Buck Woody. Introducing Microsoft SQL Server 2019: Reliability, scalability, and security both on premises and in the cloud, Packt Publishing, 2020.
5. Jiawei Han, Jian Pei & Hanghang Tong. *Data Mining: Concepts and Techniques*, Morgan Kaufmann, 4th edition, 2023.
6. Steven Hughes and Adam Jorgensen. Hands-On SQL Server 2019 Analysis Services: Design and query tabular and multi-dimensional models using Microsoft's SQL Server Analysis Services, Packt Publishing, 2020.
7. Bill H. Inmon. *Building the Data Warehouse*, John Wiley, 3e édition, 2002.
8. Ralph Kimball & Margy Ross. *Guide pratique de modélisation dimensionnelle*, Vuibert informatique, Paris, 2002.
9. Ralph Kimball & Margy Ross. *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling*, ~~Ind.,~~ Wiley, 2013

10. Stéphane Tufféry. *Data mining et statistique décisionnelle – L'intelligence des données*, éditions TECHNIP, 2012.
11. Alejandro Vaisman & Esteban Zimányi. Data Warehouse Systems - Design and Implementation, Springer, 2014.
12. Ian H. Witten, Eibe Frank, Mark A. Hall, Christopher & J. Pal. Data Mining: Practical Machine Learning Tools and Techniques, 4th Edition, Morgan Kaufmann, 2017.
13. Documentation sur RapidMiner Studio.
    https://docs.rapidminer.com/latest/studio/getting-started/ Page consultée le 5 décembre 2023
14. Module Turbo Prep de RapidMiner Studio
15. https://docs.rapidminer.com/latest/studio/guided/turbo-prep/ Page consultée le 5 décembre 2023

9. Page Web du cours :

https://moodle.uqo.ca