**ORIGINAL COURSE IMPLEMENTATION DATE:** January 2012
**REVISED COURSE IMPLEMENTATION DATE:** September 2019
**COURSE TO BE REVIEWED** (six years after UEC approval)**:** March 2025
Course outline form version: 05/18/2018

# OFFICIAL UNDERGRADUATE COURSE OUTLINE FORM

**Note: The University reserves the right to amend course outlines as needed without notice.**

| | |
|---|---|
| **Course Code and Number:** STAT 431 | **Number of Credits:** 3 Course credit policy (105) |

**Course Full Title:** Data Mining
**Course Short Title:**
*(Transcripts only display 30 characters. Departments may recommend a short title if one is needed. If left blank, one will be assigned.)*

| | |
|---|---|
| **Faculty:** Faculty of Science | **Department (or program if no department):** Mathematics & Statistics |

**Calendar Description:**

Data mining provides the techniques of extracting useful information and hidden patterns from a massive amount of data. Main topics include data exploration, classification, decision trees, Bayesian classifiers, frequent item sets, association rules, clustering, K-means, EM algorithm, and anomaly detection. Students will complete a group project based on a large real-life data set.

Note: This course is offered as STAT 431 and COMP 431. Students may take only one of these for credit.

| | |
|---|---|
| **Prerequisites (or NONE):** | COMP 230 (formerly CIS 230), STAT 271, and STAT 331/COMP 331. |
| **Corequisites (if applicable, or NONE):** | |
| **Pre/corequisites (if applicable, or NONE):** | |

**Antirequisite Courses** *(Cannot be taken for additional credit.)*
Former course code/number: **MATH 431**
Cross-listed with: **COMP 431**
Dual-listed with:
Equivalent course(s):
*(If offered in the previous five years, antirequisite course(s) will be included in the calendar description as a note that students with credit for the antirequisite course(s) cannot take this course for further credit.)*

**Typical Structure of Instructional Hours**

| | |
|---|---|
| Lecture/seminar hours | 25 |
| Tutorials/workshops | |
| Supervised laboratory hours | 25 |
| Experiential (field experience, practicum, internship, etc.) | |
| Supervised online activities | |
| Other contact hours: | |
| **Total hours** | **50** |

Labs to be scheduled independent of lecture hours: ☒ No ☐ Yes

**Special Topics** *(Double-click on boxes to select.)*
This course is offered with different topics:
☒ No ☐ Yes *(If yes, topic will be recorded when offered.)*

**Independent Study**
If offered as an Independent Study course, this course may be repeated for further credit: *(If yes, topic will be recorded.)*
☒ No ☐ Yes, repeat(s) ☐ Yes, no limit

**Transfer Credit**
Transfer credit already exists: *(See **bctransferguide.ca**.)*
☒ No ☐ Yes
Submit outline for (re)articulation:
☒ No ☐ Yes *(If yes, fill in transfer credit form.)*

**Grading System**
☒ Letter Grades ☐ Credit/No Credit

**Maximum enrolment (for information only):** 36

**Expected Frequency of Course Offerings:**
Annually *(Every semester, Fall only, annually, etc.)*

| | | |
|---|---|---|
| **Department / Program Head or Director:** Ian Affleck | **Date approved:** | November 19 2018 |
| **Faculty Council approval** | **Date approved:** | January 11, 2019 |
| **Dean/Associate VP:** | **Date approved:** | January 11, 2019 |
| **Campus-Wide Consultation (CWC)** | **Date of posting:** | February 22, 2019 |
| **Undergraduate Education Committee (UEC) approval** | **Date of meeting:** | March 1, 2019 |

**Learning Outcomes:**

Upon successful completion of this course, students will be able to:

1. Use summary statistics to describe univariate and multivariate data sets;
2. Apply different definitions of distances to measure the similarity and dissimilarity between data objects;
3. Classify discrete and continuous data using decision trees;
4. Employ different classifiers (rule-based, nearest-neighbor, naïve Bayes) to solve various classification problems;
5. Generate various frequent itemsets and set up association rules in market basket analysis;
6. Evaluate association patterns by interest factor;
7. Use the K-means clustering technique to classify data in different clusters;
8. Determine the correct number of clusters;
9. Apply the EM algorithm to cluster data by estimating the parameters of a mixture model;
10. Detect outliers in univariate and multivariate normal distributions; and
11. Complete a group project based on a large real-life data set and present the findings.

**Prior Learning Assessment and Recognition (PLAR)**

☒ Yes ☐ No, PLAR cannot be awarded for this course because

**Typical Instructional Methods** *(Guest lecturers, presentations, online instruction, field trips, etc.; may vary at department's discretion.)*

Lectures, and use of computer software. All classes take place in a computer lab.

**NOTE: The following sections may vary by instructor. Please see course syllabus available from the instructor.**

**Typical Text(s) and Resource Materials** *(If more space is required, download Supplemental Texts and Resource Materials form.)*

| | Author (surname, initials) | Title (article, book, journal, etc.) | Current ed. | Publisher | Year |
|---|---|---|---|---|---|
| **1.** | P. Tan et al. | Introduction to Data Mining | ☒ | Pearson | 2018 |
| **2.** | M. Kantardzic | Data Mining---Concepts, Models, Methods and Algorithms (2nd edition) | ☒ | Wiley-IEEE Press | 2011 |
| **3.** | | | ☐ | | |
| **4.** | | | ☐ | | |
| **5.** | | | ☐ | | |

**Required Additional Supplies and Materials** *(Software, hardware, tools, specialized clothing, etc.)*

Statistical software such as SAS Enterprise Miner.

**Typical Evaluation Methods and Weighting**

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Final exam: | 40% | Assignments: | 15% | Field experience: | % | Portfolio: | % |
| Midterm exam: | 20% | Project: | 25% | Practicum: | % | Other: | % |
| Quizzes/tests: | % | Lab work: | % | Shop work: | % | Total: | 100% |

**Details (if necessary):**

The above percentages may vary among instructors and years. The final exam is comprehensive. Students must obtain at least 40% on the final exam to pass the course.

**Typical Course Content and Topics**

**Introduction:** The origins of data mining, data mining tasks.

**Exploring data:** Types of data, data quality, data preprocessing, measures of similarity and dissimilarity, Minkowski distance, summary statistics, visualization, OLAP and multi-dimensional data analysis.

**Classification:** Decision tree induction, model over-fitting, evaluating the performance of a classifier, methods for comparing classifiers, rule-based classifiers, nearest-neighbor classifiers, Bayesian classifiers, artificial neural network, support vector machine, ensemble methods.

**Association analysis:** Frequent itemset generation, rule generation, maximal frequent itemsets, closed frequent itemsets, FP-Growth algorithm, evaluation of association patterns, interest factor, handling categorical and continuous attributes, sequential patterns.

**Cluster analysis:** Different types of clusters, K-means, bisecting K-means, agglomerative hierarchical clustering, density-based clustering algorithm, determining the correct number of clusters, supervised measures of cluster validity, EM algorithm, minimum spanning tree clustering.

**Anomaly detection:** Causes of anomalies, determining outliers in univariate and multivariate normal distributions, proximity-based outlier detection, density-based outlier detection, clustering-based techniques.