Pontificia Universidad Católica de Chile
Escuela de Ingeniería
Departamento de Ciencia de la Computación

Minería de Datos IIC2433

Ayudantía Midterm

Enunciado 1 : Integración de datos y PCA

1. Suponga que tiene una base de datos, donde el atributo Universidad tiene datos con el nombre de la Universidad completo y en otros con el nombre de la universidad abreviado, i.e. "Pontificia Universidad Católica de Chile" y "PUC". ¿Qué pasos seguiría para integrar los datos? Explique.

2. Ahora, en el mismo set de datos se encuentran los atributos "Nota", "Edad", "Año Ingreso" y "Año Egreso", además de "Universidad". En varias filas en cada uno de los atributos, indistintamente, existen datos "NaN", en otros no aparece información. ¿Qué se debe hacer con esos datos? Explique.

3. Considerando un set de datos con atributos "Universidad", "Carrera", "Nota", "Edad", "Sexo", "Año Ingreso", "Año Egreso", "Nivel Socioeconómico", "Ingresos", "Comuna". Se desea generar un algoritmo que que nos de una predicción sobre cuánto demorará un alumno en egresar, Explique detalladamente cada paso que debe seguir para que la base de datos esté preparada para el entrenamiento de dicho algoritmo. Explique cada toma de decisión y para qué sirve cada paso.

4. Si para el set de datos descritos en la pregunta 3, deseamos hacer nuestro algoritmo más eficiente, ¿qué técnica aplicamos para reducir nuestro set de datos sin perder tanta información? ¿ Explique cada paso de dicha técnica? ¿Qué atributos cree ud. serán los más relevantes? Justifique.

Enunciado 2 : FP-Growth

1. La Universidad desea generar un recomendador de cursos optativos para los alumnos de primer año a partir de los cursos que eligieron durante el primer semestre. Se tiene la información descrita en Tabla Cursos.

Alumno	**Cursos**
Alumno 1	Coro, Handball, Acuarela, Inglés
Alumno 2	Fotografía, Coro, Handball, Acuarela, Inglés
Alumno 3	Fotografía, Coro, Acuarela, Inglés
Alumno 4	Coro, Acuarela, Inglés
Alumno 5	Inglés
Alumno 6	Coro, Acuarela, Fotografía
Alumno 7	Coro, Acuarela, Handball
Alumno 8	Fotografía

Table 1: Tabla Cursos

(a) Considerando $min_{suPP} = 3$, genere el árbol con el algoritmo FP-Growth.

(b) Indique los frequent patterns generados en cada paso.

(c) Si un alumno tomó "Acuarela" en su primer semestre, ¿qué curso le recomendaría para su segundo semestre? Justifique.