Polytech Montpellier – IG4

**Data Mining**

Duration: 1:00 – Only iPad allowed

The organizers of the stage race are preparing the next seasons and would like to limit the number of racers quitting the race. The following tables have been defined to store the data over the last races.

*Race(RaceId,Year,RaceName,Organizer)*
*Racer(IdR,NameR,Age,City)*
*Team(IdT,NameT)*
*Belongs(IdR,RaceId, IdT)*
*Stage(IdSt, RaceId, DepartureCity, ArivalCity)*
*Performance(IdR,IdSt,time)*
*StageInfo (IdSt,date)*
*WeatherRecord(Date, City,temperature,windspeed,humidity)*
*WeatherType(temp,windspeed,humidity,type)*
*Distance(City1, City2,distance_km)*
*LocD (City, Department)*
*LocR(Department,Region)*
*LocC (Region, Country)*
*Terrain(City1,City2,TerrainType)*

**Question I.** What table do you propose for analyzing what impacts the number of racers quitting the race? Provide the schema of a view, the SQL query to compute it and propose some sample data (10 tuples).

**Question II.** How would a decision tree algorithm be applied on these data?

**Question III.** What is a confusion matrix? Provide an example based on the view you have proposed in Question I.

What is a false positive in your case? What are the accuracy and precision from your matrix example? Compute these measures over the sample confusion matrix.

**Question IV.** How could frequent itemsets be discovered from such data? Provide the FP-Tree and the frequent items for your sample data with a minimum support of 20%.

Polytech Montpellier – IG4

**Data Mining**

Duration: 1:00 – Only iPad allowed

The organizers of the stage race are preparing the next seasons and would like to limit the number of racers quitting the race. The following tables have been defined to store the data over the last races.

*Race(RaceId,Year,RaceName,Organizer)*
*Racer(IdR,NameR,Age,City)*
*Team(IdT,NameT)*
*Belongs(IdR,RaceId, IdT)*
*Stage(IdSt, RaceId, DepartureCity, ArivalCity)*
*Performance(IdR,IdSt,time)*
*StageInfo (IdSt,date)*
*WeatherRecord(Date, City,temperature,windspeed,humidity)*
*WeatherType(temp,windspeed,humidity,type)*
*Distance(City1, City2,distance_km)*
*LocD (City, Department)*
*LocR(Department,Region)*
*LocC (Region, Country)*
*Terrain(City1,City2,TerrainType)*

**Question I.** What table do you propose to mine for analyzing what impacts the number of racers quitting the race? Provide the schema of a view, the SQL query to compute it and propose some sample data (10 tuples).

**Question II.** How could frequent itemsets be discovered from such data? Provide the FP-Tree and the frequent items for your sample data with a minimum support of 30%.

**Question III.** How would the k-nearest neighbours algorithm be applied on these data?

**Question IV.** What is a confusion matrix? Provide an example based on the view you have proposed in Question I.

What is a false negative in your case? What are the accuracy and recall from your matrix example? Compute these measures over the sample confusion matrix.