Aplicación de Minería de Datos para el pronóstico del clima en Australia

SIGLA	NOMBRE ASIGNATURA	
MDY7101	Minería de datos	

NOMBRE			SECCIÓN	
RUT			FECHA	

PUNTAJES Y NOTA / ESCALA DE EXIGENCIA (60 %)			
PUNTAJE TOTAL: 130	NOTA: 7.0	PUNTAJE OBTENIDO	
PUNTAJE: 77	NOTA: 4.0	NOTA	

1 COMPRENSIÓN DEL NEGOCIO

En esta sección se desarrolla la comprensión del negocio, que corresponde a las primera etapa de un primer ciclo de la metodología **CRISP-DM** (**Cr**oss **I**ndustry **S**tandard **P**rocess for **D**ata Mining) para el desarrollo del proyecto de Minería de Datos aplicada al pronóstico.

1.1 Problemática Actual

Australia en los últimos 10 ha presentado uno de los mayores cambios climáticos de su historia, por lo cual se han instalado diferentes sistemas de medición climática en distintos lugares de Australia, estas observaciones son diarias y proporcionan una gran cantidad de datos.

Los gremios asociados a la agricultura del país han presentado un alto nivel de interés respecto de esta información, ya que año tras año se ven afectados por los cambios, y en la actualidad reciben esta información en formato csv, donde los datos están, pero la gran mayoría de ellos no saben cómo interpretar o cómo utilizar.

1.2 Beneficios e Hipótesis del proyecto

Anteriormente se desarrolló un proyecto OLAP, en el cual se realizó un análisis multivariado respecto de las condiciones climáticas proporcionadas por todos los centros de medición, donde se plantearon las primeras hipótesis que serán complementadas con un enfoque de minería de datos, predictivo supervisado, que pueda aportar al negocio en abordar las siguientes problemáticas:

1) **Predecir si lloverá o no al día siguiente**, dentro del conjunto de datos, se encuentra la variable **RainTomorrow** que responde a la pregunta ¿llovió al día siguiente, sí o no? Esta columna es Sí, si la lluvia de ese día fue de 1 mm o más.

2) **Conocer cuántas veces ocurrió que 3 días seguidos tuvieron lluvia en cada lugar.** Y poder analizar los mejores lugares para plantar determinadas plantas.

3) **Analizar cómo se comporta cada característica en una ubicación diferente**, considerando valores máximos y mínimos de las características en los diferentes lugares en los cuales se realizan las mediciones.

4) **Conocer cuál es la relación entre lluvia y ubicación**, tratar de responder preguntas como, ¿Cuál es la mayor cantidad de lluvia jamás registrada?, en qué ubicación llueve más días seguidos, cual es la relación entre ubicación y mm. de lluvia caída por temporadas.

2 COMPRENSIÓN DE LOS DATOS

2.1 Acerca de la fuente de datos

Los datos fueron consolidados y depurados para el análisis OLAP del primer proyecto, presenta 23 campos:

- Date: Fecha de la observación
- Location: El nombre común de la ubicación de la estación meteorológica.
- MinTemp: La temperatura mínima en grados centígrados.
- MaxTemp: La temperatura máxima en grados centígrados.
- Rainfall: La cantidad de lluvia registrada para el día en mm.
- Evaporation: Evaporación de la bandeja de clase A (mm) en las 24 horas a las 9 a.
- Sunshine: Cantidad de horas de sol brillante en el día.
- WindGustDir: Dirección de la ráfaga de viento más fuerte en 24 horas hasta la medianoche
- WindGustSpeed: Velocidad (km / h) de la ráfaga de viento más fuerte en 24 horas hasta la medianoche
- WindDir9am: Dirección del viento a las 9 am.
- WindDir3pm: Dirección del viento a las 15 pm.
- WindSpeed9am: Velocidad (km / h) del viento a las 9am.
- WindSpeed3pm: Velocidad (km / h) del viento a las 15pm.
- Humidity9am: Humedad (g/m3) del aire a las 9am. (g/m3 = gramos de agua por cada metro cúbico de aire)
- Humidity3pm: Humedad (g/m3) del aire a las 15pm. (g/m3 = gramos de agua por cada metro cúbico de aire)
- Pressure9am: Presión (pascales) del aire a las 9am.
- Pressure15pm: Presión (pascales) del aire a las 15pm.
- Cloud9am: Nubosidad (octas) del cielo a las 9am. (octas, a la parte de la bóveda celeste cubierta de nubes, hasta un máximo de 8 para el cielo cubierto o entoldado.)
- Cloud15pm: Nubosidad (octas) del cielo a las 15pm.
- Temp9am: Temperatura (C) a las 9am.
- Temp15pm: Temperatura (C) a las 15pm.
- RainToday: Si llueve más de 1mm en el día de la muestra
- RainTomorrow: predicción si lloverá más de 1 mm mañana

Fuente de datos
Las observaciones se obtuvieron de numerosas estaciones meteorológicas. Las observaciones diarias están disponibles en http://www.bom.gov.au/climate/data.

Un ejemplo de las últimas observaciones meteorológicas en Canberra:
http://www.bom.gov.au/climate/dwo/IDCJDW2801.latest.shtml

Definiciones adaptadas de http://www.bom.gov.au/climate/dwo/IDCJDW0000.shtml
Fuente de datos: http://www.bom.gov.au/climate/dwo/ y http://www.bom.gov.au/climate/data.

2.2 Variable Objetivo

La variable de supervisión o clase objetivo para este proyecto de minería de datos será el campo denominado **RainTomorrow** que responde a la pregunta ¿llovió al día siguiente, sí o no? Esta columna es Sí, si la lluvia de ese día fue de 1 mm o más.

3 PREPARACIÓN DE LOS DATOS

-

Como se indicó en la sección anterior, la fuente de datos ya se encuentra consolidada desde su origen, lo cual fue realizado para el desarrollo de análisis OLAP, por lo que no es necesario hacer un trabajo para unir las diferentes fuentes de datos. Sin embargo, al considerarse datos integrados y que serán la clave para obtener buenos resultados del pronóstico de las características climáticas, se debe realizar un proceso de limpieza y transformación de los datos a utilizar. Durante el depurado y limpieza de la base de datos se espera poder gestionar los valores perdidos y/o nulos y manejar las inconsistencias presentes y/o campos incompletos que pudieran presentarse, así como la aplicación de variables *dummies* cuando corresponda.

Se espera recibir un archivo jupyter notebook con la siguiente información:

1. Carga de datos (5 pts)
2. Análisis exploratorio inicial que incluya:
   a. Mostrar cantidad de observaciones y características (5 pts)
   b. Medidas estadísticas básicas de las columnas, dependiendo de su tipo (10 pts)
   c. Limpieza y transformación de datos a Reemplace los datos nulos o blancos según criterio, explique la decisión. (20 pts)
   d. Responda las siguientes preguntas:
      a) ¿Cuál es la diferencia de presión entre el rango mínimo y máximo que se produce a las 9 am? (5 pts)

      b) Indique las temperaturas mínimas y máximas que se producen a las 3 pm (5 pts)

      c) Indique cuantos días ha precipitado según la característica RainToday (utilice variables dummie) (20 pts)

      d) Cuantas observaciones poseen las siguientes localidades: (10 pts)
         a. Portland
         b. NorfolkIsland
         c. Moree
         d. Albany
         e. Albury

      e) Indique la localidad que posee el mayor número de observaciones (10 pts)

      f) Indique la localidad que posee el menor número de observaciones (10 pts)

   e. Proponer, al menos, 3 preguntas que puedan ser respondidas a través de agrupaciones (15 pts)

   f. Mostrar algunas tendencias, distribuciones de los datos. Deberá usar al menos 3 gráficos para mostrar las tendencias/distribuciones. Cada gráfico debe incluir un análisis de lo que muestra. (15 pts)