Tecnológico de Costa Rica - Fundatec

Programa Ciencia de los Datos

Minería de Datos e Inteligencia de Negocio

Profesora Lorena Zúñiga

Estudiante Sergio Castillo Segura

Avance 1 Proyecto Final

Predicción de lluvia en Australia

10 marzo 2020

Contenido

Entendimiento de Negocio

Objetivos de Negocio

Se intenta generar un método de predicción que permite estimar la probabilidad de lluvia 24 horas antes de que suceda con el fin de mantener informada a la población con un muy buen porcentaje de exactitud y se puedan tomar previsiones para las actividades al aire libre.

Criterios de Éxito

El método generado deberá poder predecir si lloverá o no al día siguiente con un porcentaje de exactitud del 90%.

Situación Actual

Inventario de Recursos

Se cuenta con una base de datos que contiene 142000 muestras de datos meteorológicos recolectados día tras día desde el 1 de Noviembre 2007, hasta el 25 de junio 2017 en 40 distintas localidades de Australia. Las estaciones meteorológicas recolectan datos a lo largo del día por lo que algunas columnas son agregadas con valores máximos y mínimos, pero además hay datos precisos obtenidos 2 veces al día, a las 9am y a las 3pm, los cuales permiten medir la evolución del clima durante el día.

Entre los datos se incluye temperatura, cantidad de lluvia en mm, evaporación, brillo del sol, viento, humedad, presión atmosférica, etc.

Requerimientos

- Se requiere generar un modelo predictivo que permita determina con un 90% de precisión la presencia de lluvia dentro de las siguientes 24 horas de acuerdo a los datos meteorológicos tomados durante el día y proyectando los registros diarios obtenidos durante los últimos 10 años.
- El modelo debe ser capaz de predecir lluvia para cualquiera de las 40 localidades presentes en el dataset provisto.
- El modelo también debería poder predecir lluvia en los siguientes 3 días con un porcentaje de precisión del 70%.

Supuestos

Se deberá suponer que la data contenida en el dataset es correcta y suficiente para poder hacer el modelo predictivo y fue recolectada por fuentes oficiales de instituciones meteorológicas de Australia.

A la vez se estima que la data recolectada es válida para hacer predicciones futuras dentro de los próximos 10 años a la última fecha de datos en el dataset sin tener que hacer cambios o ajustes por cambios en el comportamiento del clima.

Restricciones

El modelo solo podrá ser usado para la predicción de lluvia en las localidades para las cuales se contiene data en el dataset original utilizado. Se entiende que el modelo no dará predicciones asertivas para ninguna otra localidad.

El dataset ya incluye una variable para cada día de la muestra que indica la cantidad de lluvia real obtenida al día siguiente, esto a través de la data histórica. Esta variable debería ser eliminada del estudio pues contiene la respuesta que se trata de buscar en el dataset a través de una medida diferente. Al ser este un problema de clasificación no se usa esta columna como criterio de éxito, pero si se quisiera se podría utilizar para generar un modelo de regresión que trate de predecir de forma más precisa no solo si lloverá o no sino la cantidad aproximada.

Riesgos

Numero	Riesgo	Contingencia	Prioridad
1	Puede que la cantidad de lluvia en Australia sea muy baja lo que limite la usabilidad del modelo.	Se deberá asumir el riesgo como parte del estudio.	Baja
2	Puede que existan datos faltantes o inconsistentes que ensucien las muestras produciendo errores en el modelo.	Se deberá hacer un buen estudio preliminar de los datos junto con una limpieza y normalización del dataset.	Alta
3	Puede que la agregación de las medidas diarias no brinde el nivel de granularidad suficiente para obtener el resultado esperado.	Se deberá asumir el riesgo pues no existe mayor granularidad de los datos.	Baja
4	Puede que el modelo resulte ser muy técnico para cierto tipo de usuarios lo que limite su usabilidad.	Se deberá genera un modelo sencillo de utilizar basado en fechas para predicciones futuras.	Media

Beneficios

Debido a que la probabilidad de lluvia en Australia es realmente baja, muchas veces la lluvia toma desprevenida a la población quienes preparan eventos y actividades al aire libre, así como labores de cosecha y riego en zonas agrícolas, que si no se toman las medidas preventivas necesarias podrían incurrir en afectaciones económicas importantes.

Un modelo predictivo con un alto índice de efectividad ayudaría a las autoridades meteorológicas servir mejor a la población al brindarles información altamente confiable.

Además se podría ahorrar al uso indebido de reservas de agua potable para riego y pastoreo en zonas agrícolas si se previenen esas actividades debido a la probabilidad de lluvia.

En general, un modelo predictivo de lluvia puede traer muchos beneficios para la planeación agrícola y urbana cuando es correctamente aplicado.

Plan de Proyecto

Fase	Tiempo	Recursos	Riesgos
Entendimiento de negocio	2 días	Científico de datos Dataset	Ninguno
Entendimiento de los datos	2 días	Científico de datos Dataset	Ninguno
Preparación de los datos	2 días	Científico de datos Dataset R Studio	Ninguno
Modelado	2 días	Científico de datos Dataset limpio R Studio	Puede que el modelo elegido no converja y se tome más tiempo en la prueba y elección de otros modelos.
Evaluación	1 día	Científico de datos Dataset de pruebas R Studio Modelo de datos	Puede que el modelo creado no tenga un buen rendimiento con los datos de evaluación y se tenga que volver a la fase de modelado.
Implementación	2 días	Científico de datos R Studio Modelo de datos Python	Se asumirá que la versión final del modelo será una herramienta web generada en python, aunque no será el objetivo de este estudio.

Evaluación de herramientas y técnicas

Se realizará todo el trabajo de preparación, limpieza, transformación, modelaje y evaluación utilizando el lenguaje R y las diferentes bibliotecas disponibles.

Se generarán tres modelos de clasificación binaria utilizando Regresión Logística, KNN y SVM, y se elegirá el que tenga mejor resultado entre los 3.

Objetivos de minería de datos

Crear un modelo de clasificación binaria que permite predecir si habrá lluvia o no en las siguientes 24 horas para 40 localidades en Australia con un porcentaje de exactitud de al menos 90%.

Criterios de éxito

El modelo será exitoso si logra predecir con exactitud si lloverá o no al día siguiente de la toma de datos meteorológicos, la precisión es binaria en términos de llueve o no llueve. Un resultado opcional será predecir la cantidad de lluvia dentro de un margen de error de +-2mm de lluvia.

Entendimiento de los datos

Conjunto de datos requeridos

Se requiere un dataset que contenga las medidas meteorológicas de las 40 localidades Australianas que se quieren incluir en el modelo predictivo, a razón de un registro por día por localidad para los últimos 10 años.

Método de acceso

El dataset será provisto en un archivo de texto con formato CSV con valores numéricos y textuales, para efectos de este estudio el dataset obtenido es estático, por lo que si existen nuevos registros el modelo generado deberá ser vuelto a entrenar para incluirlos.

Descripción de los datos

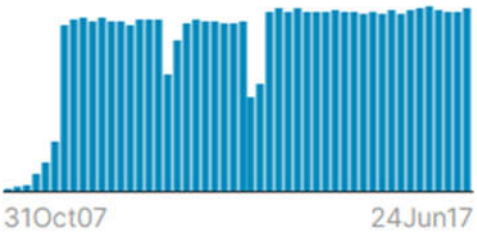
A continuación una descripción de cada columna contenida en el dataset a utilizar:

Columna	Descripción
Fecha (Date)	Indica la fecha de la muestra en formato [m]m/[d]d/yyyy
Locación (Location)	Nombre de la localidad donde fue tomada la muestra.
Temp. Mínima (MinTemp)	Indica el valor mínimo de la temperatura durante el día en grados centígrados.
Temp. Máxima (MaxTemp)	Indica el valor máximo de la temperatura durante el día en grados centígrados.
Lluvia caída (Rainfall)	La cantidad de lluvia caída durante el día en mm.
Evaporación	La evaporación plena de las últimas 24 horas en mm.
Sol (Sunshine)	El número de horas que el sol brillo en el cielo con visibilidad plena.
Direccion del viento (WindGustDir)	La dirección de la máxima aceleración del viento durante las utlimas 24 horas.
Velocidad del viento (WindGustSpeed)	La velocidad de la máxima aceleración de viento de las últimas 24 horas.
Dirección del viento 9am (WindDir9am)	La dirección del viento a las 9am.
Dirección del viento 3pm(WindDir3pm)	La dirección del viento a las 3pm.
Velocidad del viento 9am (WindDir9am)	La velocidad del viento a las 9am.
Velocidad del viento 3pm(WindDir3pm)	La velocidad del viento a las 3pm.
Humedad 9am (Humidity9am)	Porcentaje de humedad a las 9am.
Humedad 3pm (Humidity3pm)	Porcentaje de humedad a las 3pm.

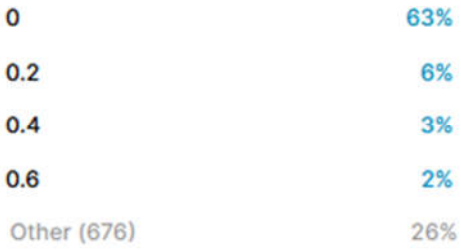
Presión 9am (Pressure9am)	Presión atmosférica (hpa) reducida al nivel del mar a las 9am.
Presión 3pm (Pressure3pm)	Presión atmosférica (hpa) reducida al nivel del mar a las 3pm.
Nubes 9am (Cloud9am)	Cantidad de octavas del cielo que son tapadas por las nubes a las 9am.
Nubes 3pm (Cloud3pm)	Cantidad de octavas del cielo que son tapadas por las nubes a las 3pm.
Temperatura 9am (Temp9am)	Temperatura en grados centígrados a las 9am.
Lluvia hoy (RainToday)	Indicación booleana sobre si llovió o no.
Riesgo de lluvia en mm (Risk_MM)	Riesgo de lluvia caída al día siguiente en mm.
Lluvia mañana (RainTomorrow)	Variable objetivo, llueve mañana si o no?

Exploración de Datos

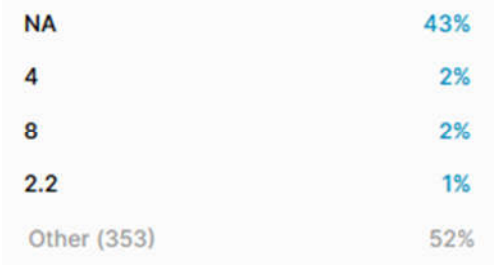
Histograma de registros



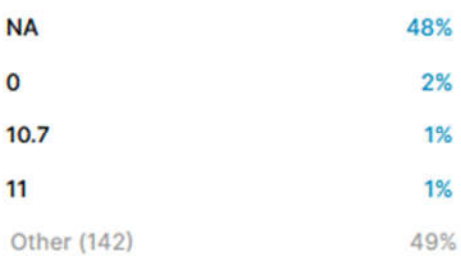
Cantidad de lluvia caída (mm)



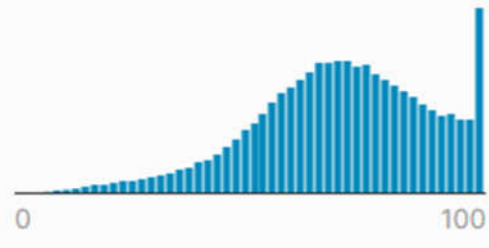
Evaporación (mm)



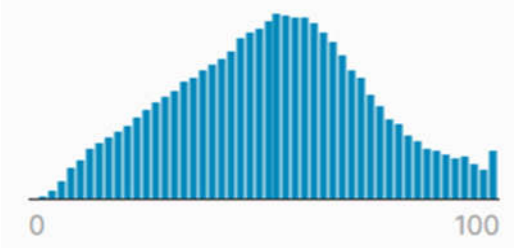
Sol



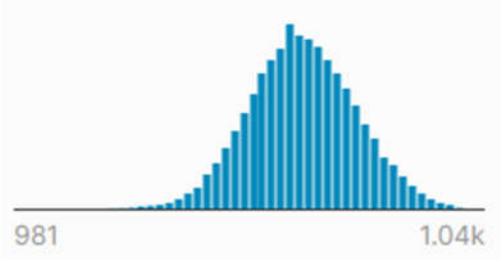
Porcentaje Humedad 9am



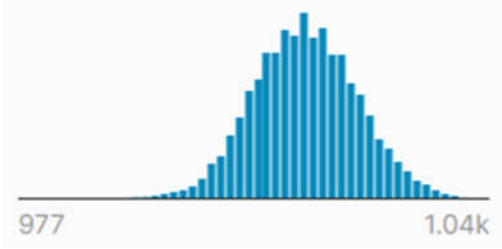
Porcentaje Humedad 3pm



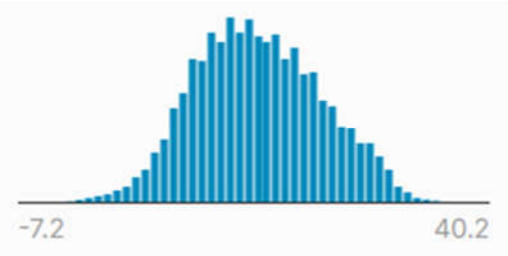
Presión Atmosférica 9am (hpa)



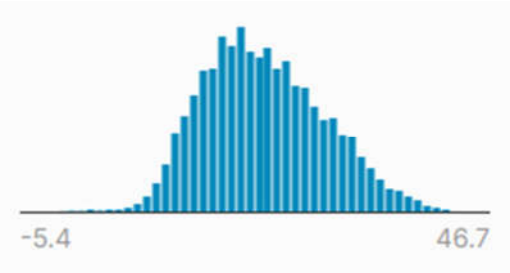
Presión Atmosférica 3pm (hpa)



Temperatura 9am



Temperatura 3pm



Calidad de datos

Existe una muy buena cantidad de registros por día que permiten alimentar a un modelo predictivo de una manera óptima, casi todas las columnas contienen datos considerados importantes para el estudio y la cantidad de datos faltantes es mínima, se puede realizar una limpieza justa sin afectar la calidad de los datos.

La calidad de los datos es óptima para el estudio.