**PARCIAL N° 2**

**MINERÍA DE DATOS**

Integrantes:

* Juan Palma
* Bastián Barahona
* Manuel Godoi

Contenido

[Introducción 3](#_Toc106007857)

[Modelos de regresión 4](#_Toc106007858)

[Modelo Lineal Múltiple 4](#_Toc106007859)

[Hipótesis 4](#_Toc106007860)

[Evaluación de Score 4](#_Toc106007861)

[Gráfico/Screenshot de código 4](#_Toc106007862)

[Modelo de Árbol de decisión de regresión 6](#_Toc106007863)

[Hipótesis 6](#_Toc106007864)

[Gráfico/Screenshot de código 6](#_Toc106007865)

[Modelos de Clasificación 8](#_Toc106007866)

[Modelo Regresión Logística 8](#_Toc106007867)

[Hipótesis 8](#_Toc106007868)

[Evaluación de Score 8](#_Toc106007869)

[Gráfico/Screenshot de código 9](#_Toc106007870)

[Modelo K-Means 13](#_Toc106007871)

[Hipótesis 13](#_Toc106007872)

[Evaluación de Score 13](#_Toc106007873)

[Gráfico/Screenshot de código 13](#_Toc106007874)

[Conclusión del modelo 16](#_Toc106007875)

[Conclusión 17](#_Toc106007876)

[Juan Palma 17](#_Toc106007877)

[Bastian Barahona 17](#_Toc106007878)

[Manuel Godoi 17](#_Toc106007879)

[Referencias 18](#_Toc106007880)

# Introducción

En este documento presentaremos la implementación de modelos de Regresión y Clasificación para analizar distintas hipótesis acerca de la caída de precipitaciones en Australia.

Diseñamos distintos modelos predictivos con el uso de algoritmos de aprendizaje automático a partir de un Dataset que contiene años de observaciones meteorológicas diarias de muchos lugares en Australia.

Específicamente la investigación se basó en el entrenamiento y testeo de dos modelos de regresión, uno perteneciente al modelo de regresión lineal múltiple y otro de Árbol de decisión de regresión. Por otro lado, también se evidencian dos modelos de clasificación, los cuales son utilizando K-Means y regresión logística.

# Modelos de regresión

## Modelo Lineal Múltiple

### Hipótesis

Gracias a la temperatura registrada a las 9 am, la velocidad del viento a las 9 am y la humedad a las 9 am, podemos predecir la temperatura máxima registrada en ese día en la locación de Albury.

### Evaluación de Score

El resultado del conjunto de datos seleccionados nos da como resultado lo siguiente:

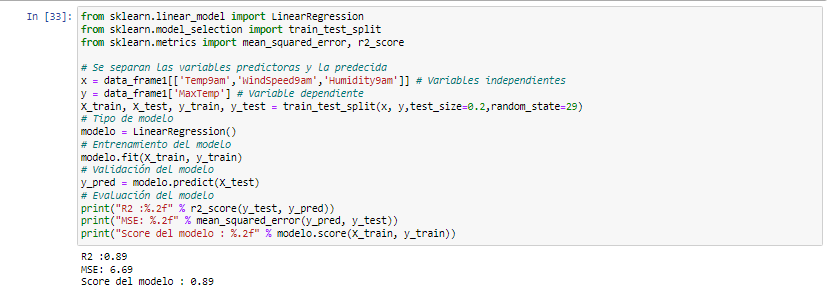
**R2 :0.89**

**MSE: 6.69**

**Score del modelo: 0.89**

Por lo cual, la predicción tiene una probabilidad de un 89% de ser acertada.

### Gráfico/Screenshot de código







Conclusión del modelo

Gracias a la data, podemos concluir que el modelo predictivo es en un 89% correcto, por lo cual, en 9 de cada diez predicciones, el modelo estaría en lo correcto o muy aproximado. Esta predicción podría llegar a ser mejor si la data analizada estuviese mejor registrada, ya que, se encontraron muchas variables con datos nulos, lo cual afecta directamente a la predicción.

### Modelo de Árbol de decisión de regresión

### Hipótesis

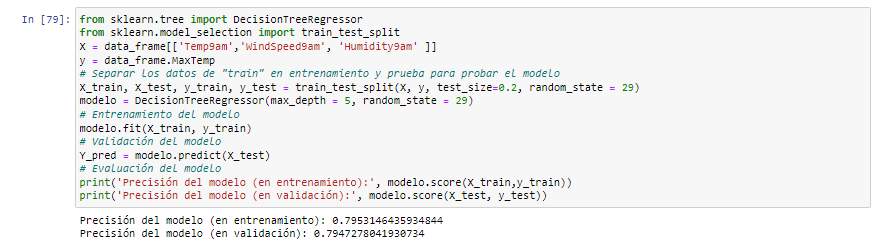
Se predice la temperatura máxima del día considerando las localidades de Australia basándose en las variables de temperatura a las 9 am, velocidad del viento a las 9 am y la humedad del entorno a las 9 am.

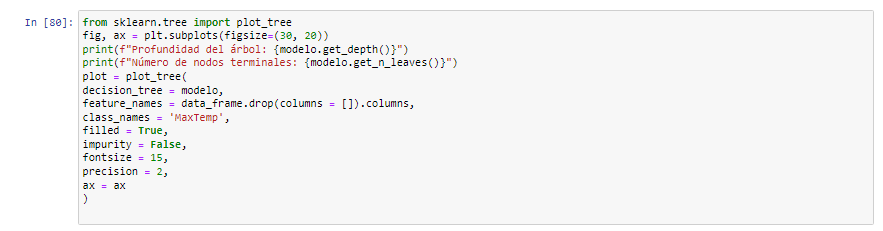
Evaluación de Score

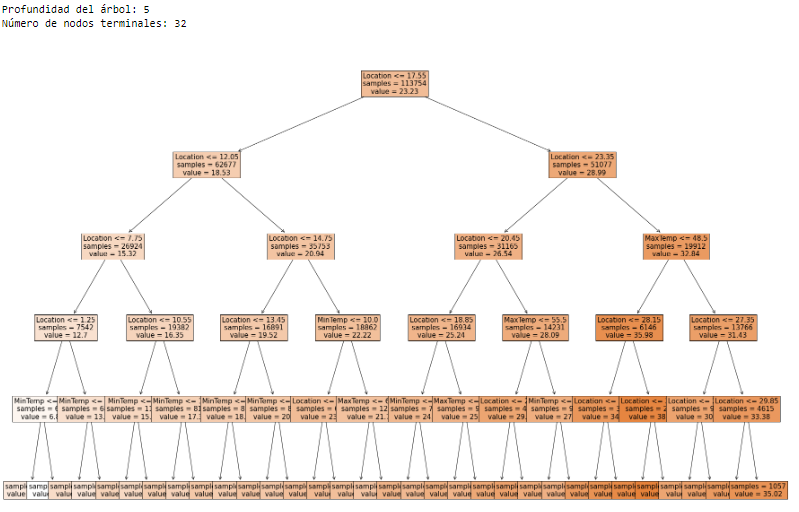
Precisión del modelo (en entrenamiento): 0.7953146435934844

Precisión del modelo (en validación): 0.7947278041930734

### Gráfico/Screenshot de código







Conclusión del modelo

Gracias al modelo de árbol de decisión de regresión, los modelos evidencian de manera muy sencilla cómo funciona el modelo de predicción. Este modelo separa por cada variable agrupando los resultados para luego volver a dividirlos hasta llegar mediante un árbol a un grupo bastante pequeño siguiendo una línea clara y sencilla de seguir

.

# Modelos de Clasificación

## Modelo Regresión Logística

### Hipótesis

El rendimiento del modelo depende de las características que se utilizan para entrenar dicho modelo. Se asignará una puntuación a cada característica de entrada en función de lo útil que será para predecir la variable Target y se considerarán los valores de mayor importancia.

### Evaluación de Score

La puntuación de Precisión del modelo de Regresión Logística se considera que hace un muy buen trabajo de predicción dado su score y el tiempo que tomó entrenar el modelo.

**Puntuación de Precisión: 0.8454235380990893**

El modelo generaliza bien los datos invisibles, ya que no muestra problemas de Overfitting o Underfitting.

**Train Data Score: 0.8422385146895933**

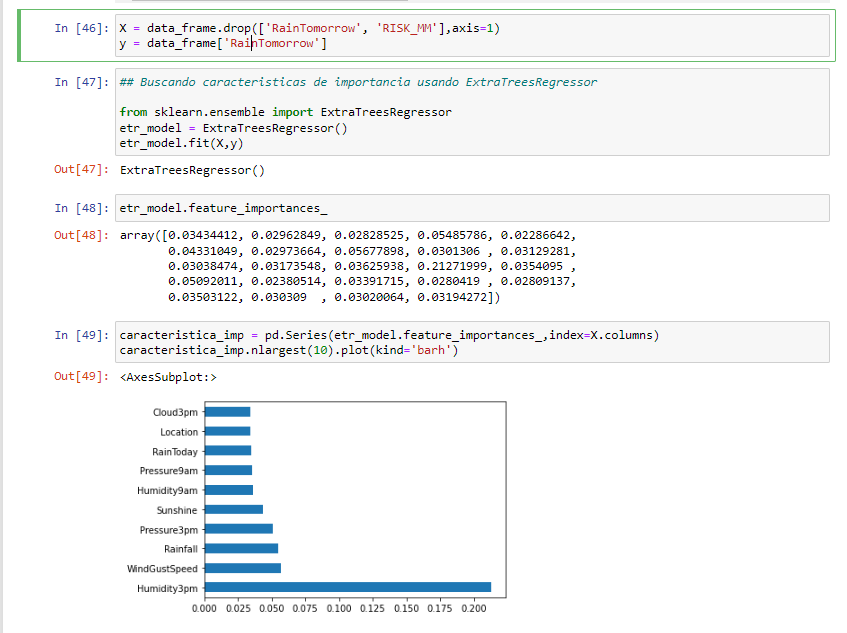
**Test Data Score: 0.8454235380990893**

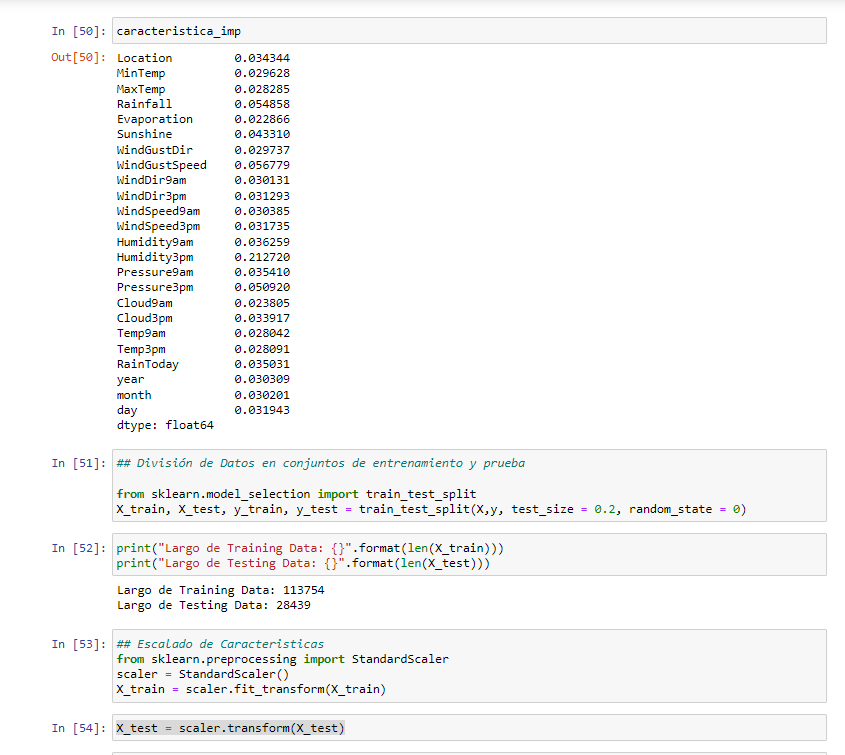
La puntuación promedio de Validación Cruzada es muy similar a la puntuación de precisión del modelo original. En consecuencia, es posible que la precisión del modelo no se mejore mediante validación cruzada.

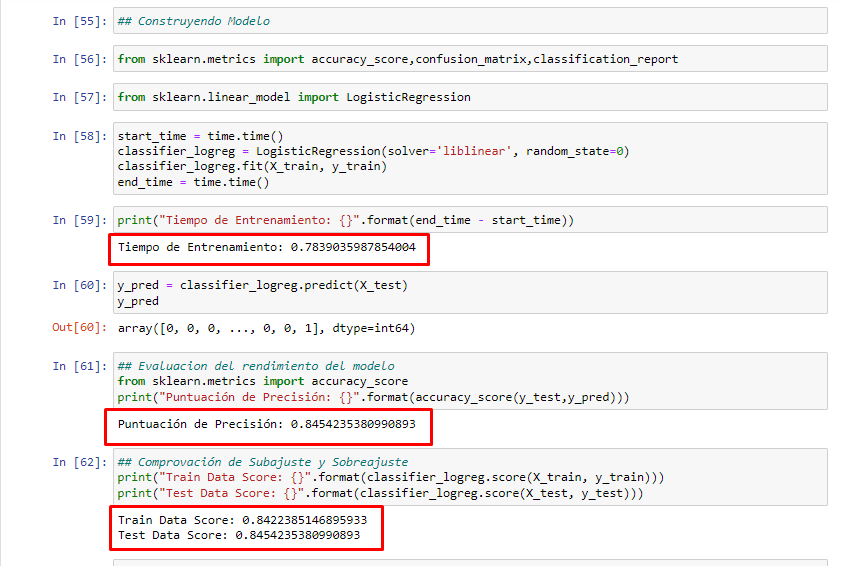
**Puntuaciones validación cruzada: [0.84325964 - 0.84031471 - 0.84005099 - 0.84536943 - 0.84232967]**

**Puntaje promedio de validación cruzada: 0.8422648879580705**

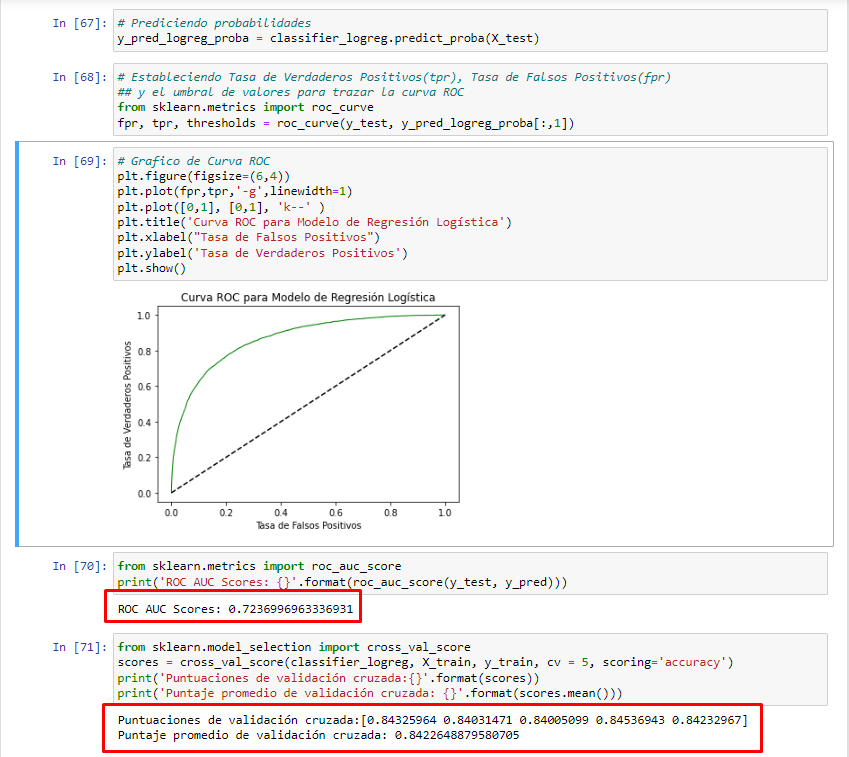
### Gráfico/Screenshot de código











## Modelo K-Means

### Hipótesis

En este caso seleccionamos 3 dimensiones: La temperatura a las 9 am ("Temp9am") , la velocidad del viento a las 9am ("WindSpeed9am"), y la humedad a las 9 am ("Humidity9am") y se cruzan para ver si entregan alguna pista de su agrupación y la relación con sus categorías (max temperatura).

### Evaluación de Score

El resultado del conjunto de datos seleccionados nos da como resultado lo siguiente:

R2 :0.80

MSE: 9.84

Score del modelo: 0.80

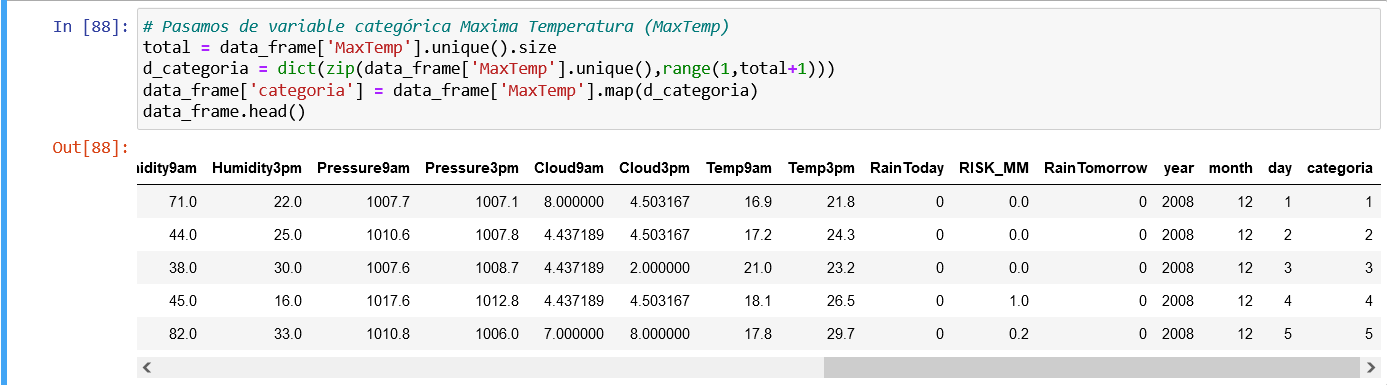
Por lo cual, la predicción tiene una probabilidad de un 80% de ser acertada.

### Gráfico/Screenshot de código

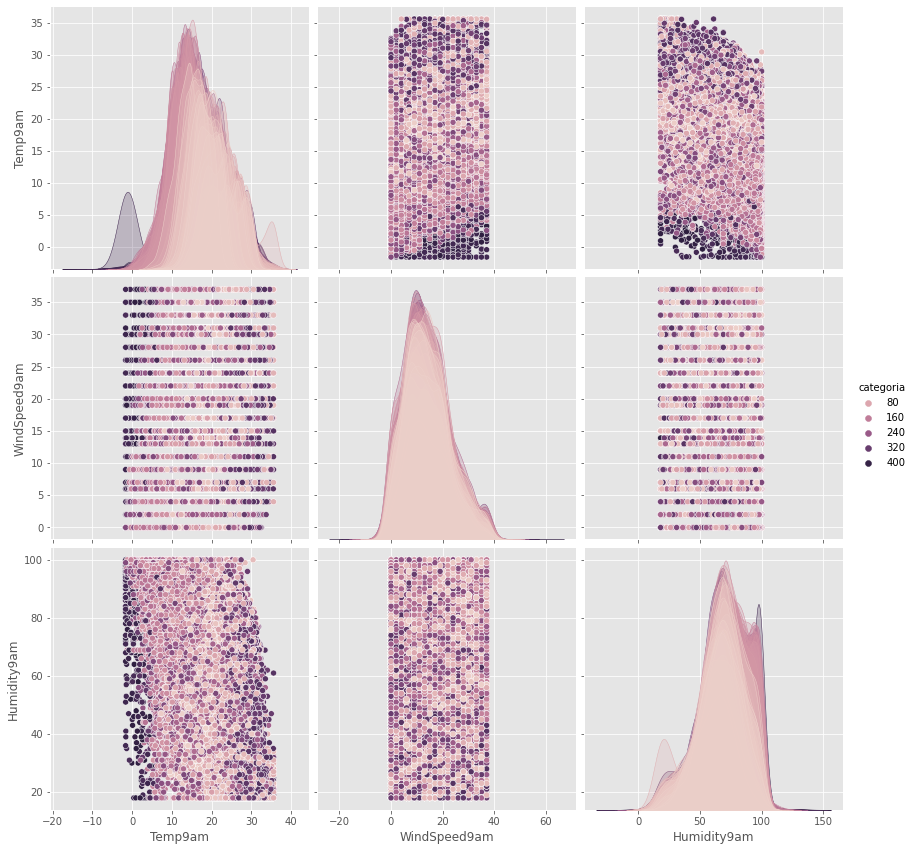




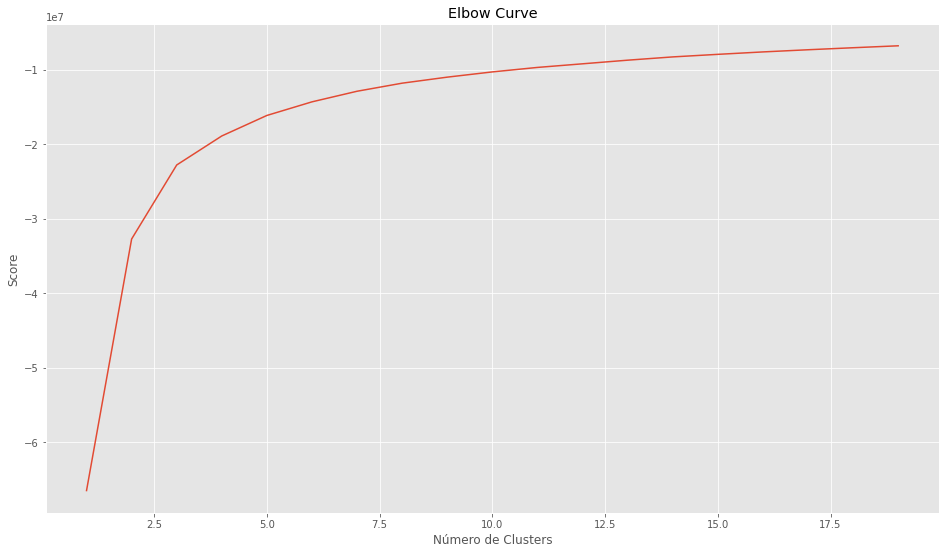








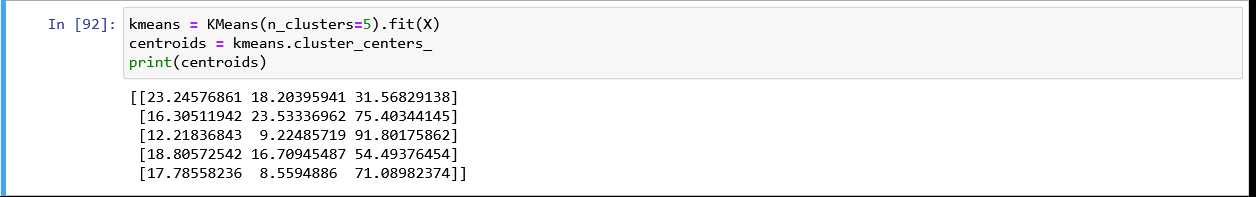




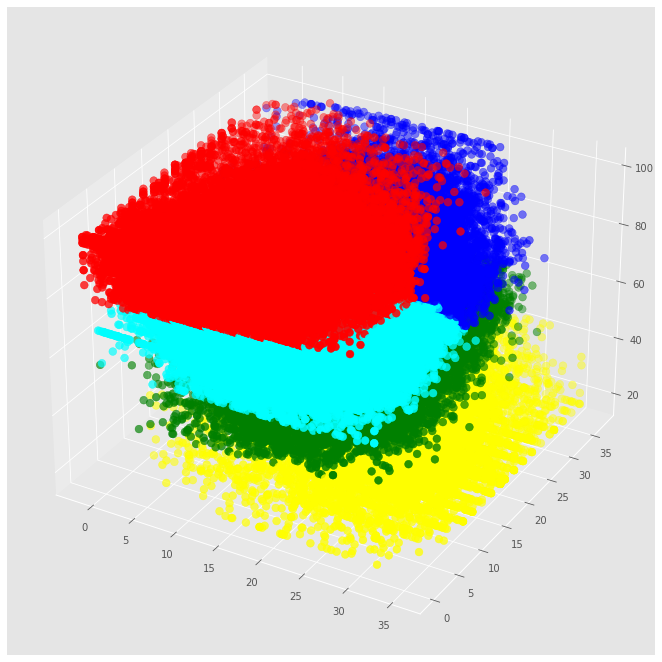
El gráfico anterior representa la curva de Elbow. Se trata de una heurística para determinar el

número de conglomerados en un conjunto de datos.

Se observa que la curva es bastante “suave”. Se va a considerar a 5 como un buen número para K.







Es posible ver que el Algoritmo de K-Means con K=5 ha agrupado a las temperaturas máximas, teniendo en cuenta las 3 dimensiones que se utilizaron: la temperatura a las 9am, la velocidad del viento a las 9am y la humedad a las 9am.

# Conclusión del modelo

El algoritmo de K-means nos ayudó a crear clúster de producto en grandes grupos de datos que no estaban etiquetados, para intentar descubrir nuevas relaciones entre los factores como lo pudimos ver con las variables ya indicadas (temperatura máxima "Maxtemp", temperatura a las 9am "Temp9am", velocidad del viento a las 9 am "WindSpeed9am" y humedad a las 9am "Humidity9am").

# Conclusión

## Juan Palma

Las herramientas que nos otorga Jupyther para trabajar son realmente poderosas, con un buen data set se pueden sacar cosas bastante interesantes, predecir no es una ciencia exacta, pero puede darnos una probabilidad bastante certera de la realidad todo gracias a la data histórica de lo que se está estudiando. Lo visto en esta unidad me hizo aprender mucho sobre cómo trabajar los datos, filtrar, procesar, predecir, etc.

## Bastián Barahona

Dado el planteamiento del problema, fue imposible no preguntarse ‘Será este el ajuste correcto’, o si es que ‘Debo mirar más profundo’. Los datos a menudo están desordenados, incompletos o son redundantes y el desarrollo de este proyecto generó un interés enorme por aprender más acerca del ciclo de vida de la Ciencia de Datos y las técnicas que se utilizan para la alimentación de un modelo de aprendizaje automático. Este trabajo no solo consta de analizar modelos o procesar variables y características, si no que invita a interactuar con la información de manera que podamos darle sentido a una investigación y valor a nuestros hallazgos.

## Manuel Godoi

En conclusión, lo pasado en clases y lo que nos muestra la información entregada pudimos evidenciar y poder hacer una forma más efectiva de poder clasificar y utilizar regresiones con el fin de obtener mejores predicciones y evidenciar ciertos patrones o parámetros según el dataset en este caso observaciones meteorológicas diarias de muchos lugares en Australia.

# Referencias

Na8 (12 diciembre de 2017). Qué es overfitting y underfitting y cómo solucionarlo. <https://www.aprendemachinelearning.com/que-es-overfitting-y-underfitting-y-como-solucionarlo>

Jose Martinez Heras (21 septiembre de 2020). Regresión Logística para Clasificación <https://www.iartificial.net/regresion-logistica-para-clasificacion/#Curiosidades>

Jose Martinez Heras (29 septiembre de 2020). ¿Clasificación o Regresión? <https://www.iartificial.net/clasificacion-o-regresion>

Max (2022). Diferencia entre clasificación y regresión en machine learning <https://ingenierobeta.com/clasificacion-vs-regresion-machine-learning>

Akshisaxena (2018). CATBOOST [https://es.acervolima.com/catboost-ml](https://ingenierobeta.com/clasificacion-vs-regresion-machine-learning)

Paloma Recuero(17 abril 2018). Python para todos (5): Termina tu primer experimento de Machine Learning con Python <https://empresas.blogthinkbig.com/python-para-todos-5-fin-del-experimento>

Natalia Acevedo (23 octubre 2020). Matriz de confusión en Machine Learning <https://nataliaacevedo.com/matriz-de-confusion-en-machine-learning-explicado-paso-a-paso>

Jose Martinez Heras (19 septiembre de 2020). Árboles de Decisión con ejemplos en Python <https://www.iartificial.net/arboles-de-decision-con-ejemplos-en-python>

Jose Martinez Heras (26 junio de 2018). Unsupervised Learning (5) - Clustering, Dimensionality Reduction <https://github.com/jmartinezheras/2018-MachineLearning-Lectures-ESA/blob/master/5_UnsupervisedLearning/5_Unsupervised_DowJones.ipynb>