**EXAMEN TRANSVERSAL DE APLICACIÓN DE MINERÍA DE DATOS PARA EL ANÁLISIS PREDICTIVO DE EVENTOS CLIMÁTICOS EN AUSTRALIA.**

Equipo : Valentina Silva – Jonathan Navarro – Ignacio Jara

Carrera : Ingeniería en Informática

Asignatura : Minería de Datos

Sección : 002D

Docente : Jazna Meza Hidalgo

Concepción, 21 de Julio de 2021

# **Tabla de contenido**

1-[Descripción del contexto 3](#_heading=h.gjdgxs)

2-[Propósito y justificación del proyecto](#_heading=h.tpw3g45vn5j) 3

3-[Análisis de datos disponibles y patrones de comportamientos.](#_heading=h.3z1in3ws89ci) 3

3.1- Exploración Inicial 3

3.2- Visualización 9

3.3- Machine Learning 15

4-[Razonamiento y justificación respecto al análisis de datos 27](#_heading=h.gtp5kua9t0cv)

5-[Definición de fases, tareas y entregables a partir del conjunto de datos y proyecto en general 28](#_heading=h.ltyyi9ykpp7v)

6-[Sugerir decisiones estratégicas 2](#_heading=h.nxu6ldn5hz74)9

# 

# 

# 

# 

# **1- Descripción del contexto**

En Australia existen varias localidades en las que el clima varía día a día, gracias a esto se pueden conseguir datos que pueden ayudar a predecir el clima o incluso buscar los mejores lugares donde instalar una granja eólica. Con este documento se pretende mostrar esto de forma visual.

# **2- Propósito y justificación del proyecto**

El propósito de este documento y proyecto es mostrar de una forma gráfica los datos seleccionados para las visualizaciones, predicciones, clasificaciones y segmentaciones, intentando conseguir así datos útiles para el futuro.

# **3- Análisis de datos disponibles y patrones de comportamientos.**

**3.1- Exploración Inicial:**

## 



Para el análisis de los datos, hemos utilizado la librería Pandas, Numpy y Matplotlib.

* Pandas se utiliza para operaciones y manipulaciones de datos estructurados. Nos entrega diferentes herramientas, con las que nos permitirá realizar la limpieza y preparación de la estructura de datos.
* Numpy nos agrega apoyo para todos los cálculos matemáticos, como la utilización de matrices, listas, entre otros.
* La librería de Matplotlib la utilizamos para expresar la información de la estructura de datos a través de gráficos, para comprender su comportamiento general o específico, según sea el caso.

Para empezar el proceso de análisis, hacemos una exploración inicial de los datos y queremos decir cual es la cantidad de observaciones y características contiene este conjunto de datos



Podemos decir hay un total de 142193 filas/observaciones con una cantidad

de 24 columnas/características cada una

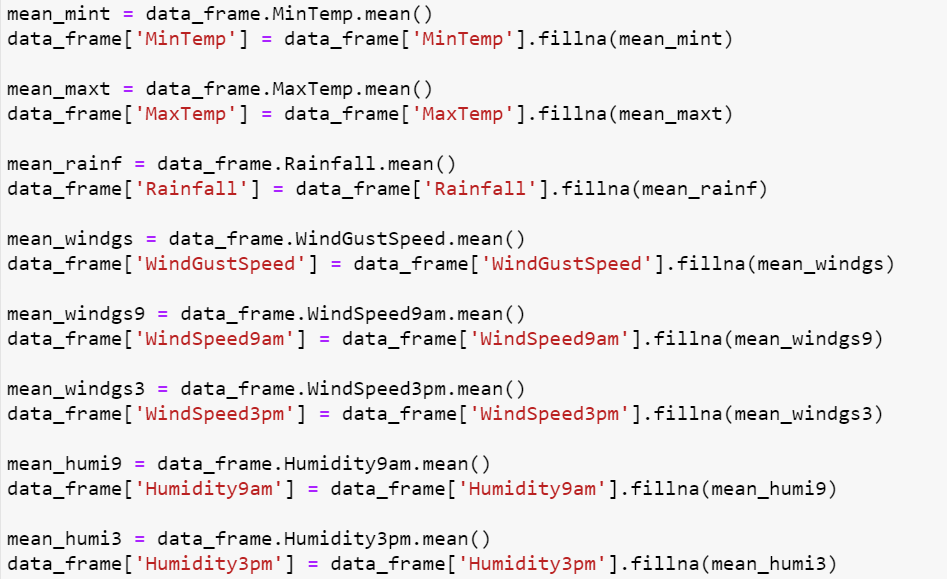
Identificar si hay existencia de datos nulos

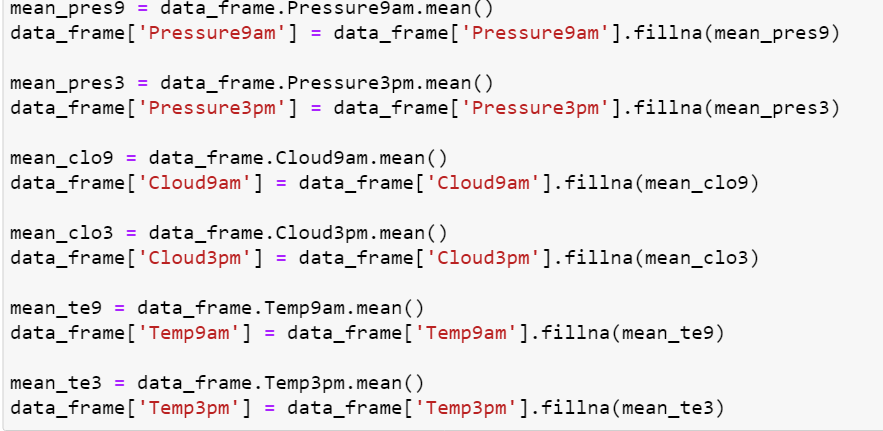


Previamente en el resultado arrojó que hay existencia de datos nulos , ahora tendrá que ver cuántos datos nulos hay por cada columna

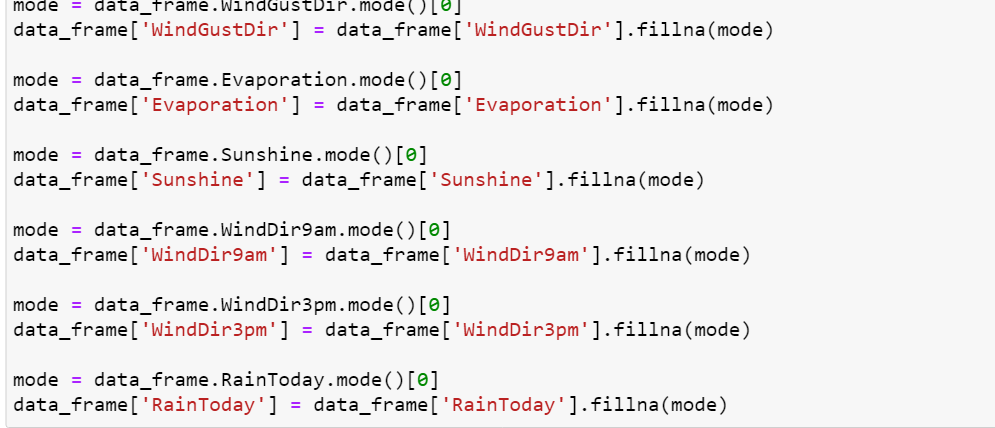


Luego de obtener la información , realizamos una limpieza de datos de valores nulos cuantitativos





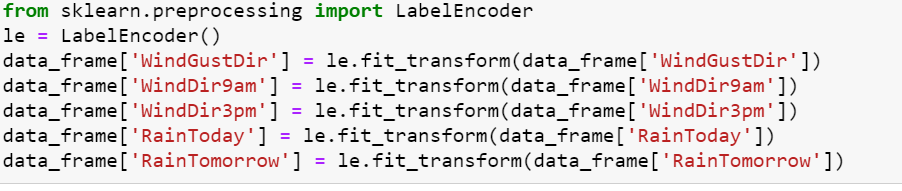
Luego de obtener esa información, ahora continuamos limpiando los datos de valores nulos cualitativos



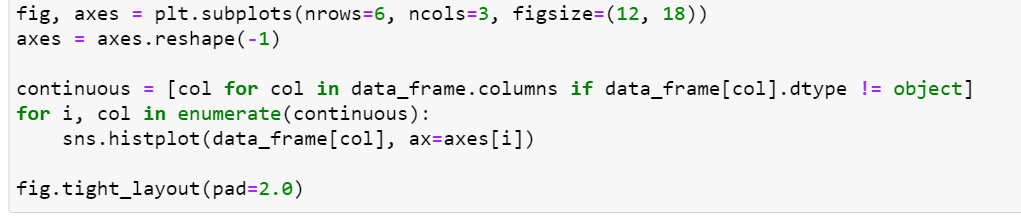
Una vez ya limpiados todos los datos, debemos revisar nuevamente si se han limpiado todos los datos nulos

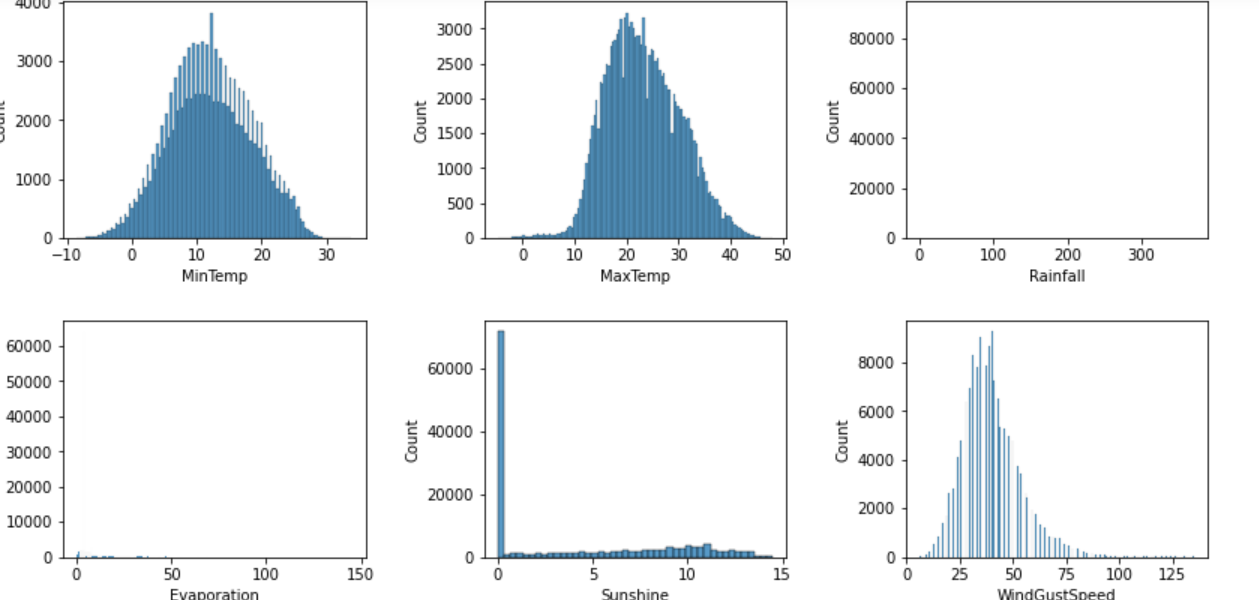


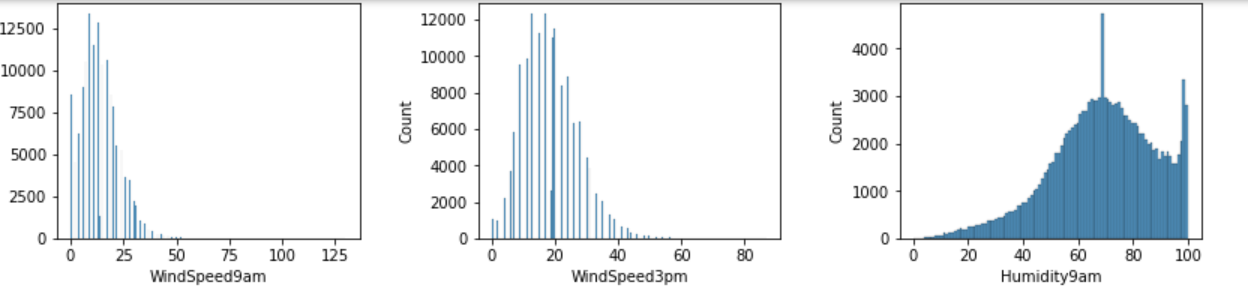
Realizamos un preprocesamiento de datos que realizamos en los datos brutos para transformarlos en datos y que tengan formatos que sean más fáciles de utilizar

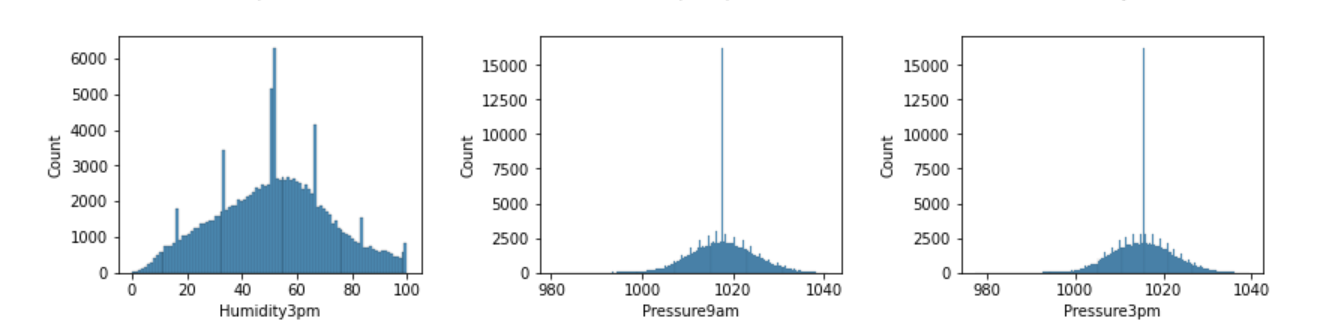


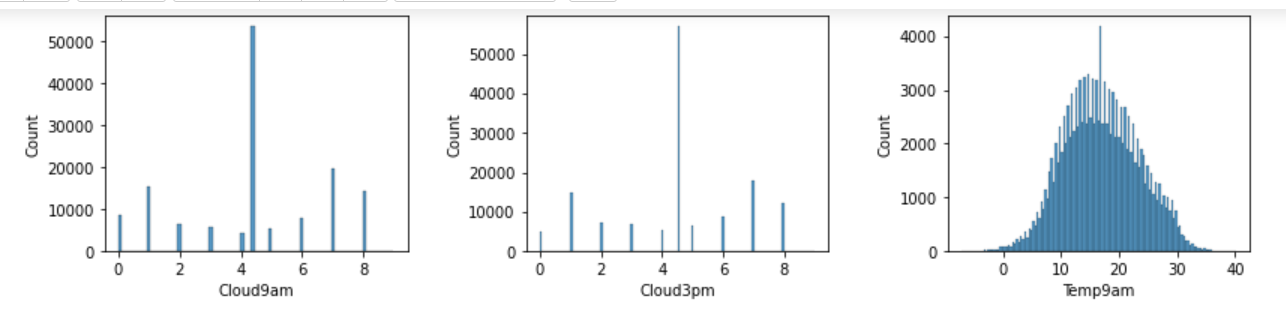
Aquí se puede ver el comportamiento de los datos dentro del data frame, como se puede apreciar la mayoría de columnas tienen datos normales a excepción de Rainfall y RISK\_MM que tiene datos anómalos o directamente no contienen datos lo cual no refleja la real situación en el dataframe

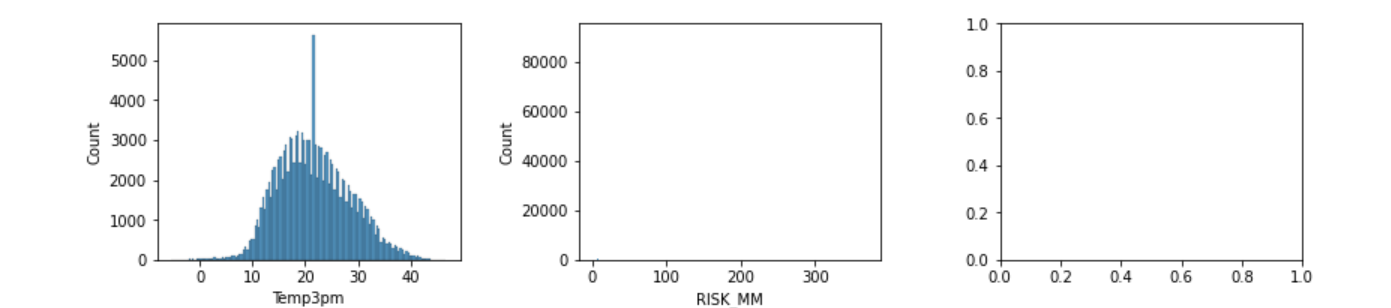




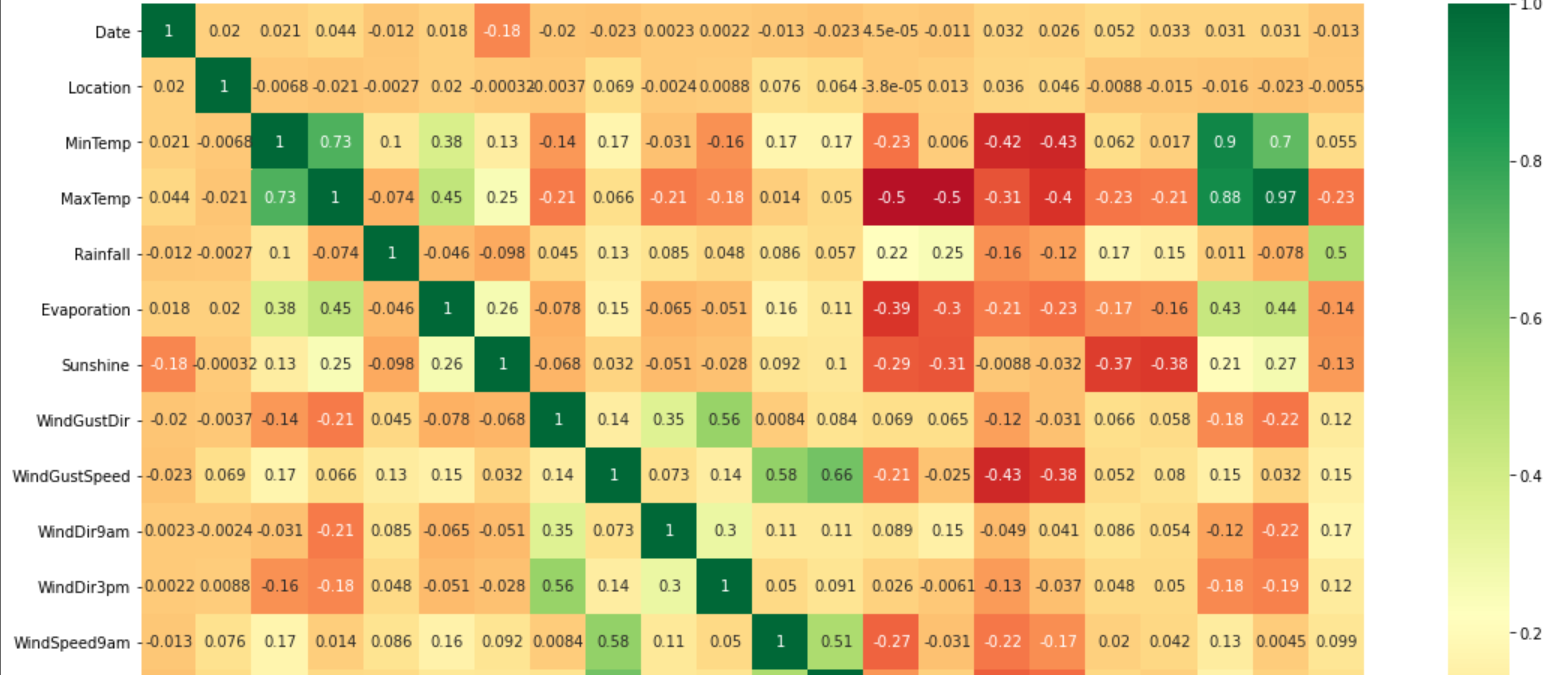


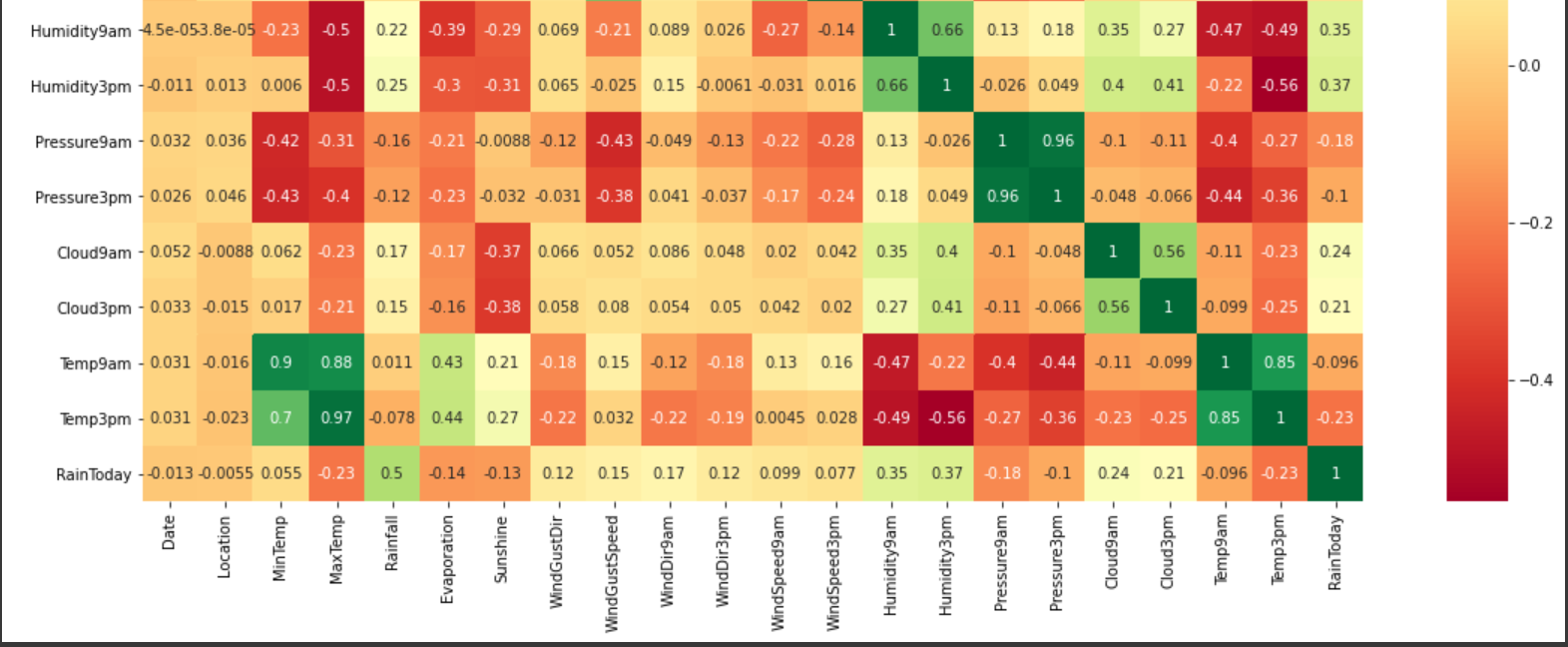






Mapa de calor

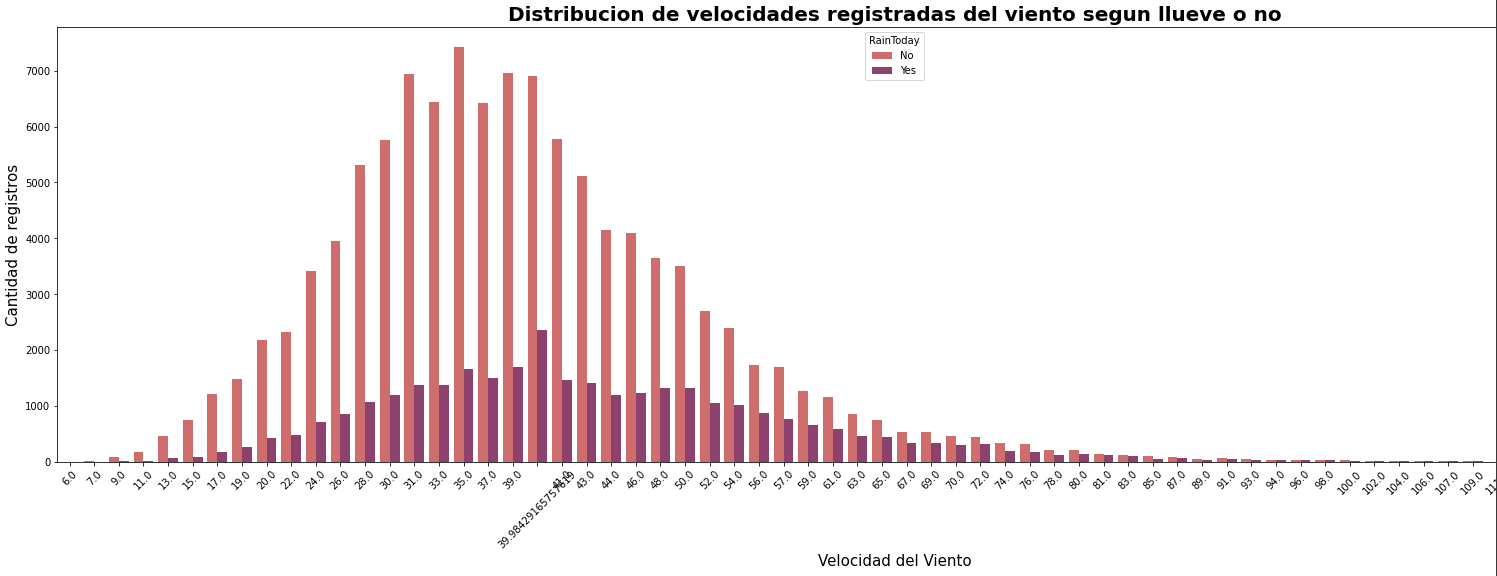




## **3.2- Visualización:**

## 

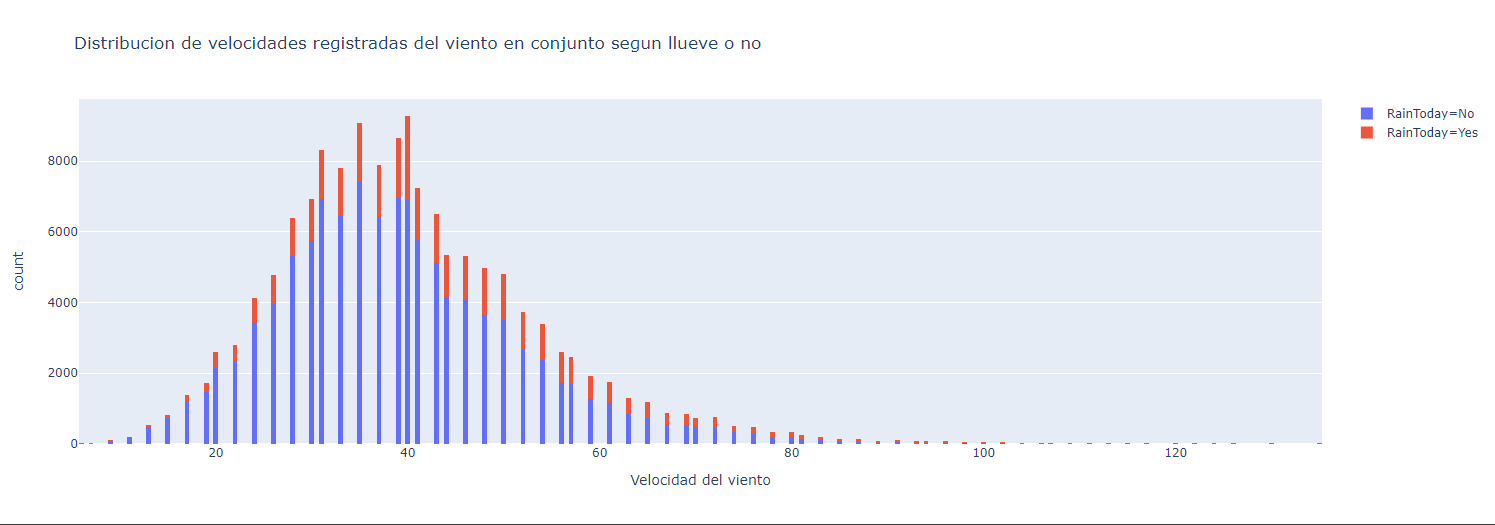
**Comparación:**



**Interpretación:**

El gráfico de barras agrupadas nos permite saber cuales son las velocidades de viento con más registros, además de poder ver la distribución de los registros de las rafagas cuando hay días lluviosos y cuando no.

Podemos inferir que hay más registros cuando no llueve, además podemos inferir que la velocidad con más registros es de 35km/h cuando no llueve y en comparación, cuando llueve la velocidad con mayor número de registros es de 39,9km/h.



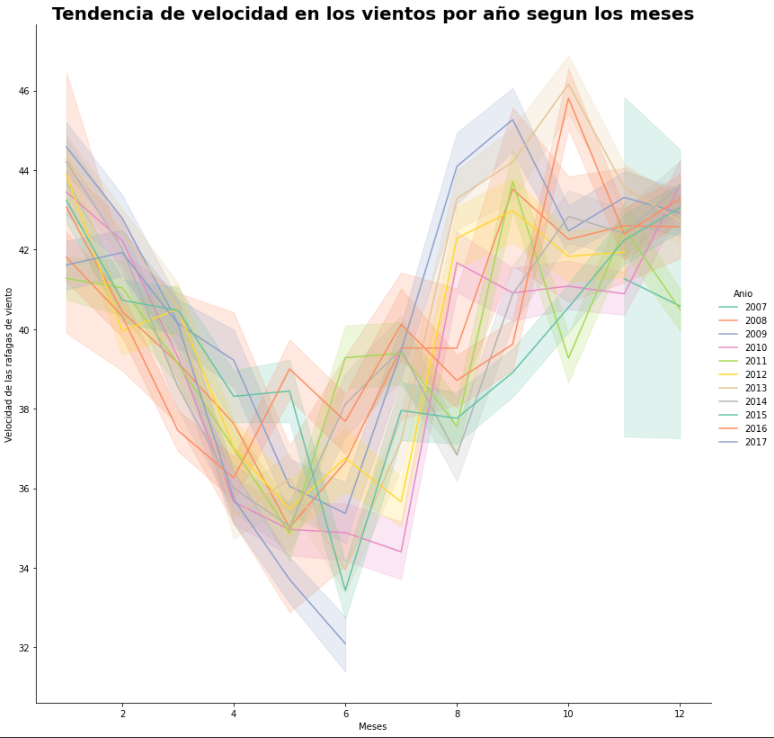
**Interpretación:**

Este histograma nos permite comparar las velocidades de viento según el pronóstico de lluvia para el mismo día.

Como se puede apreciar la velocidad que tiene más registros en los días lluviosos es de 41 km/h en cambio su contraparte sin lluvia es 35km/h indicando que por lo general en los días de lluvia tiene a haber velocidades de rafagas mucho más altas.

Esto podría servir para generar un reporte de riesgos ya que a más velocidad de viento es más probable que ocurran accidentes.

**Tendencia:**

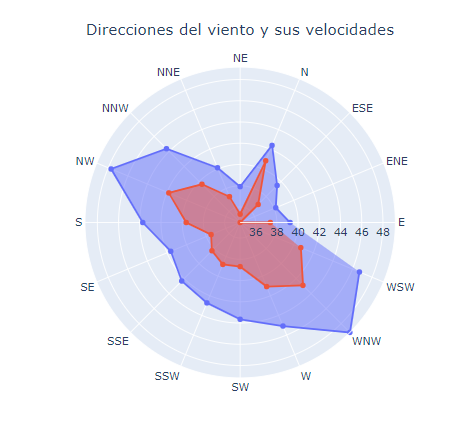
****

**Interpretación:**

Con este gráfico podemos observar la tendencias de la velocidad del viento en las distintas épocas del año.

Podemos inferir que en el año la velocidad del viento tiende a aumentar en las temporadas cálidas y tiene la tendencia de disminuir en las temporadas frías.

Ésta información puede ser de utilidad para determinar en qué meses es mejor, por ejemplo, usar los vientos para secar la ropa o incluso en qué meses es más conveniente hacer uso de la energía eólica

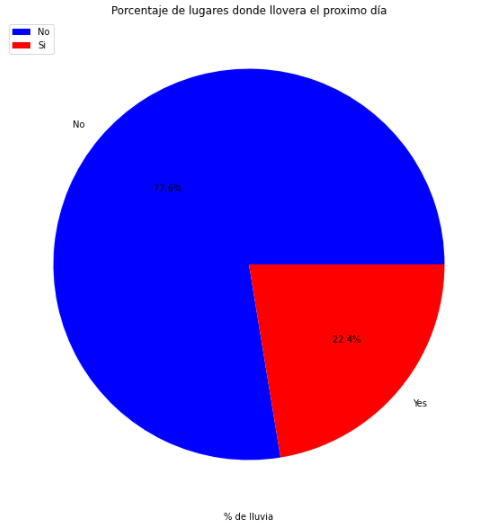
****

**Interpretación:**

Gráfico de Radar que nos permite entender el comportamiento del viento, ya que podemos ver hacia que direccion las rafagas de viento son mas fuertes.

Se puede observar en el gráfico, las rafagas mas altas de viento son hacia la dirección NorOeste y Oeste NorOeste de Australia.

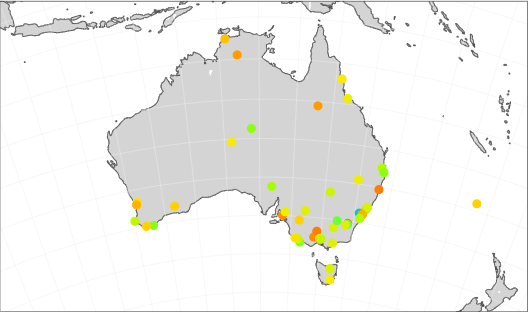
**Partes de un total:**

****

**Interpretación:**

Con este gráfico de partes de un total se busca mostrar de forma visual cual es el porcentaje de localidades donde lloverá el próximo día. Como se puede apreciar en un 22.4% de localidades lloverá el próximo día. Esta información puede servir para informar en las localidades donde lloverá sobre los resguardos que deben tener para impedir cualquier accidente.

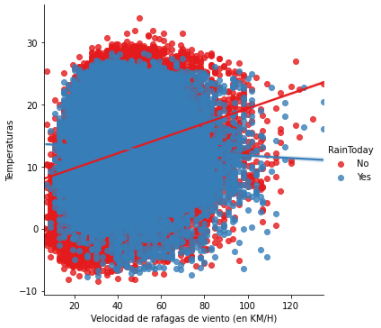
**Distribución:**

****

**Interpretación:**

El Mapa de coropletas nos permite observar geograficamente en que localidades se dan los registros de rafagas de viento mas altas, identificándose con su respectivo color dependiendo de la velocidad de la rafaga. Como se puede apreciar el lugar donde las velocidades de rafaga alcanzan la mayor velocidad es Penrith, gracias a su gran capacidad para generar rafagas de 80km/h se podría crear una granja eólica en esa localidad para generar energía a las localidades cercanas.

**Correlación:**

****

**Interpretación:**

Como se puede apreciar en el gráfico de dispersión, existe una relación entre las temperaturas mínimas y las rafagas de viento en los días con y sin lluvia.

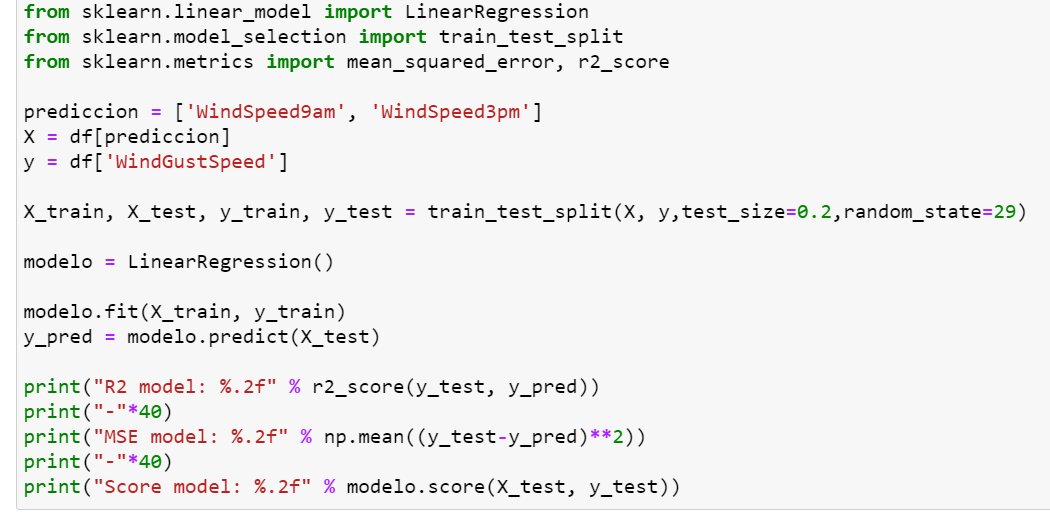
Se puede ver que en los días lluviosos las temperaturas no sobrepasan los 26°C pero tienden a tener una mayor velocidad de rafagas por el contrario si no hay lluvia las temperaturas puede llegar sobre los 30°C pero con rafagas de viento menores.

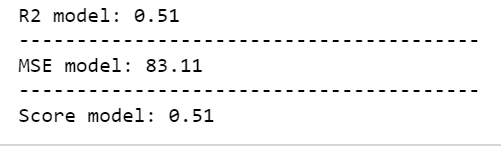
**3.3- Machine learning:**

**Predicción:**

**Regresión lineal (algoritmo ML)**

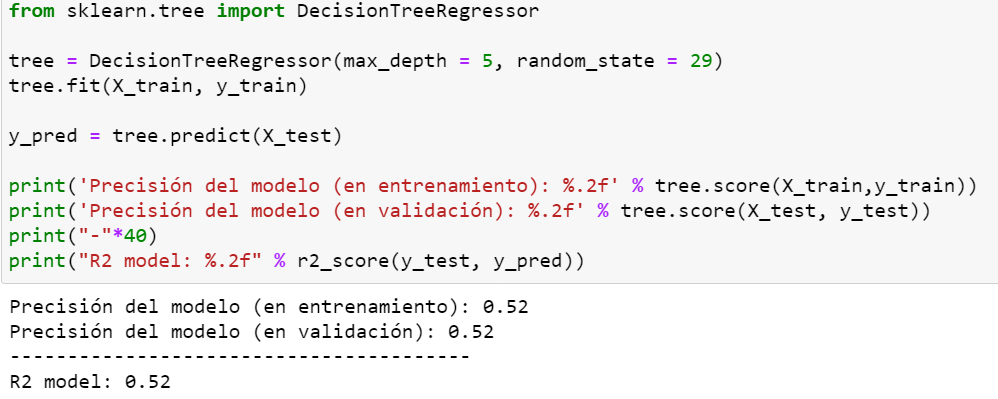
* En este modelo de predicción se usaron los datos WindSpeed9am y WindSpeed3pm para predecir el dato WindGustSpeed, estos datos se usaron con la finalidad de conseguir la mayor probabilidad de una predicción acertada cuando se quiere predecir la velocidad de las ráfagas de viento en el día según las velocidades que se alcanzaron a las 9am y las 3pm. Esto se podría utilizar en conjunto con un mapa de regiones para buscar el mejor lugar donde se puede instaurar energía eólica de forma eficiente.
* Como se puede apreciar, la certeza del modelo es del 51% siendo relativamente alta dejando un margen de error del 49%.



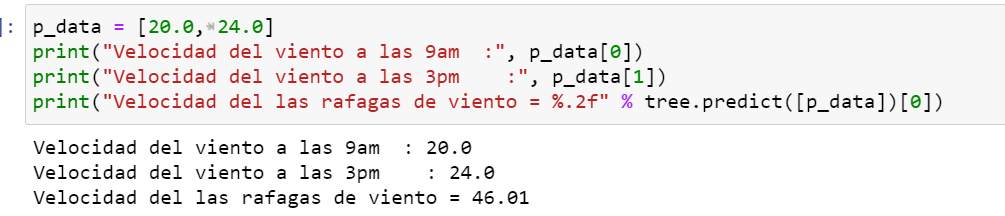


**DecisionTreeRegressor (algoritmo ML)**

El modelo de decisiontreeregressor alcanza solo un 1% más que el modelo linearregression llegando a un 52% de certeza en la predicción, haciendo que éste modelo sea el más apto para realizar la predicción

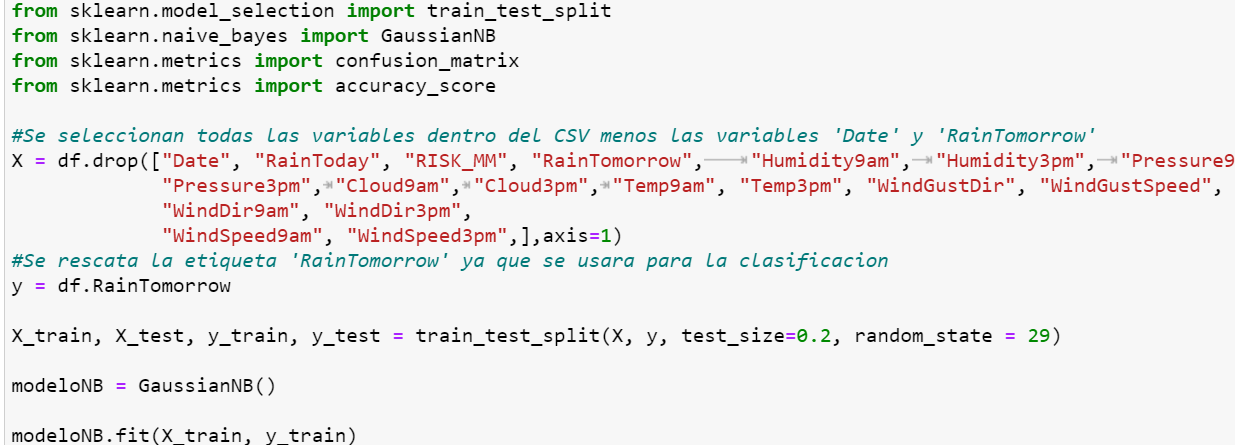


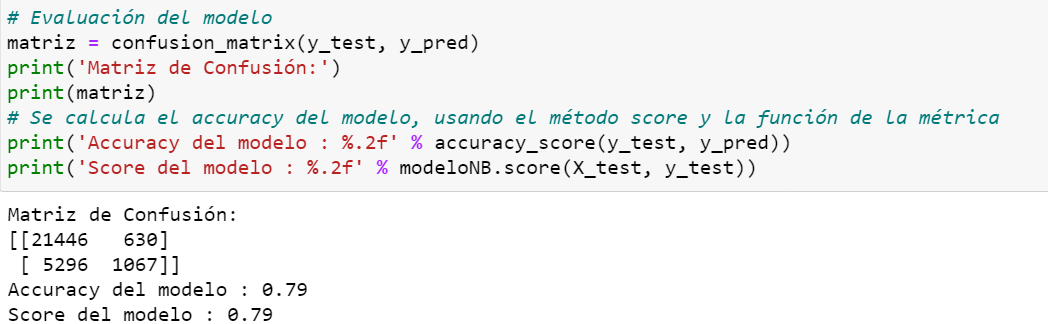
### **Predicción con el algoritmo decisiontreeregressor**



**Clasificación:**

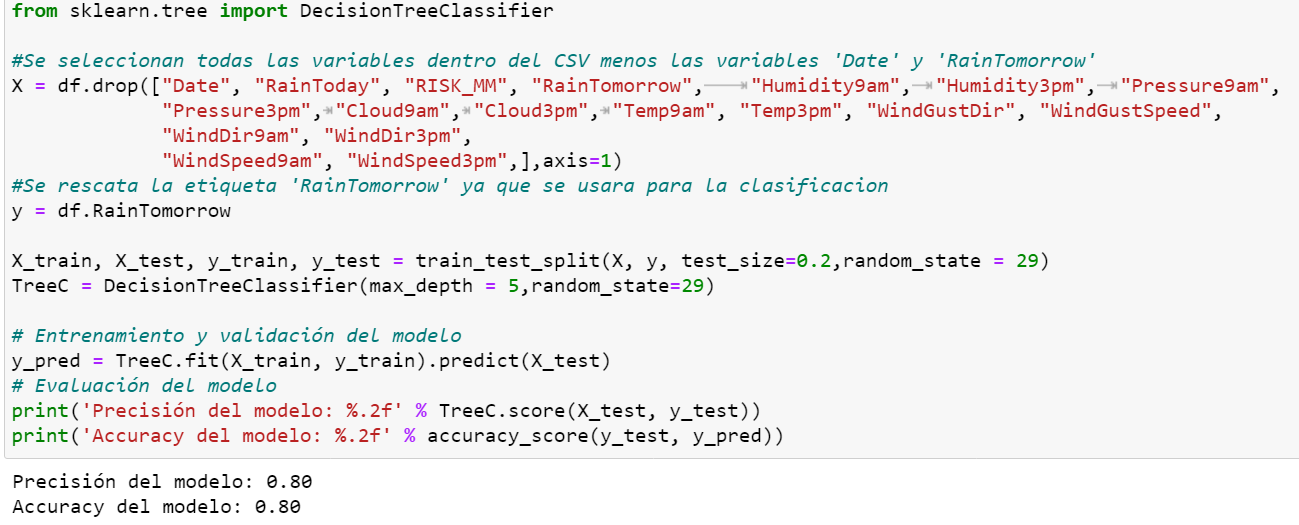
**Naive bayes (algoritmo ML)**

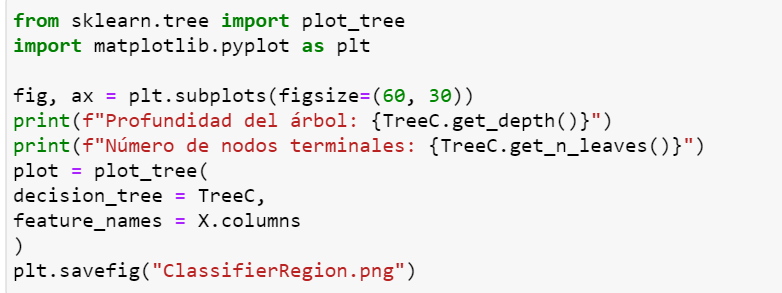
Con el algoritmo NaiveBayes se logra determinar en que lugar llovera segun las variables Location, MinTemp, MaxTemp, Rainfall, Evaporation y Sunshine que se pueden encontrar dentro del dataframe, esta clasificación tiene un 79% de certeza haciéndolo excelente para su uso.



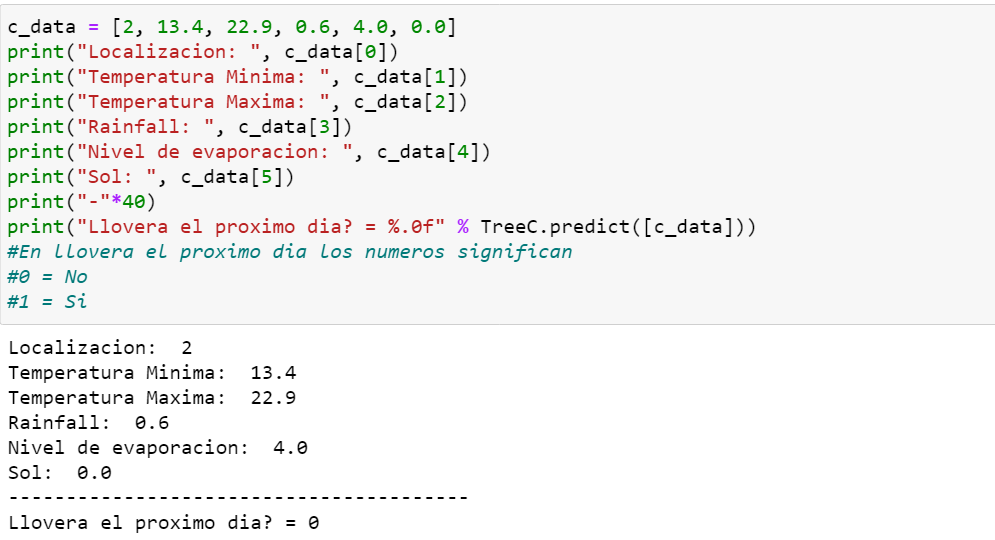
**Decisiontreeclassifer (algoritmo ML)**

Con el algoritmo DecisionTreeClassifier se logra un 80% de certeza en la clasificación haciéndolo el mejor algoritmo para utilizar, aunque desde un punto de vista técnico 1% de diferencia no afectaría mucho en los resultados.



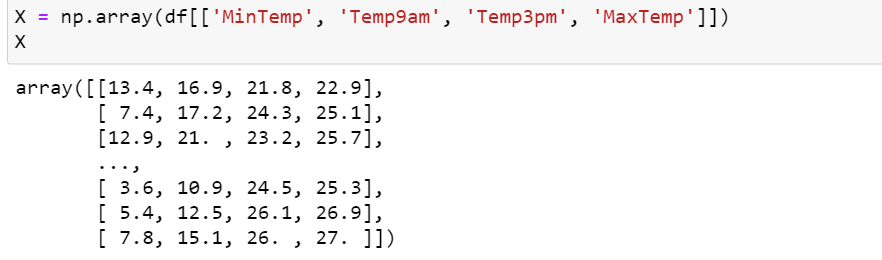
 

Uso de Decisiontreeclasifier (Algoritmo ML)

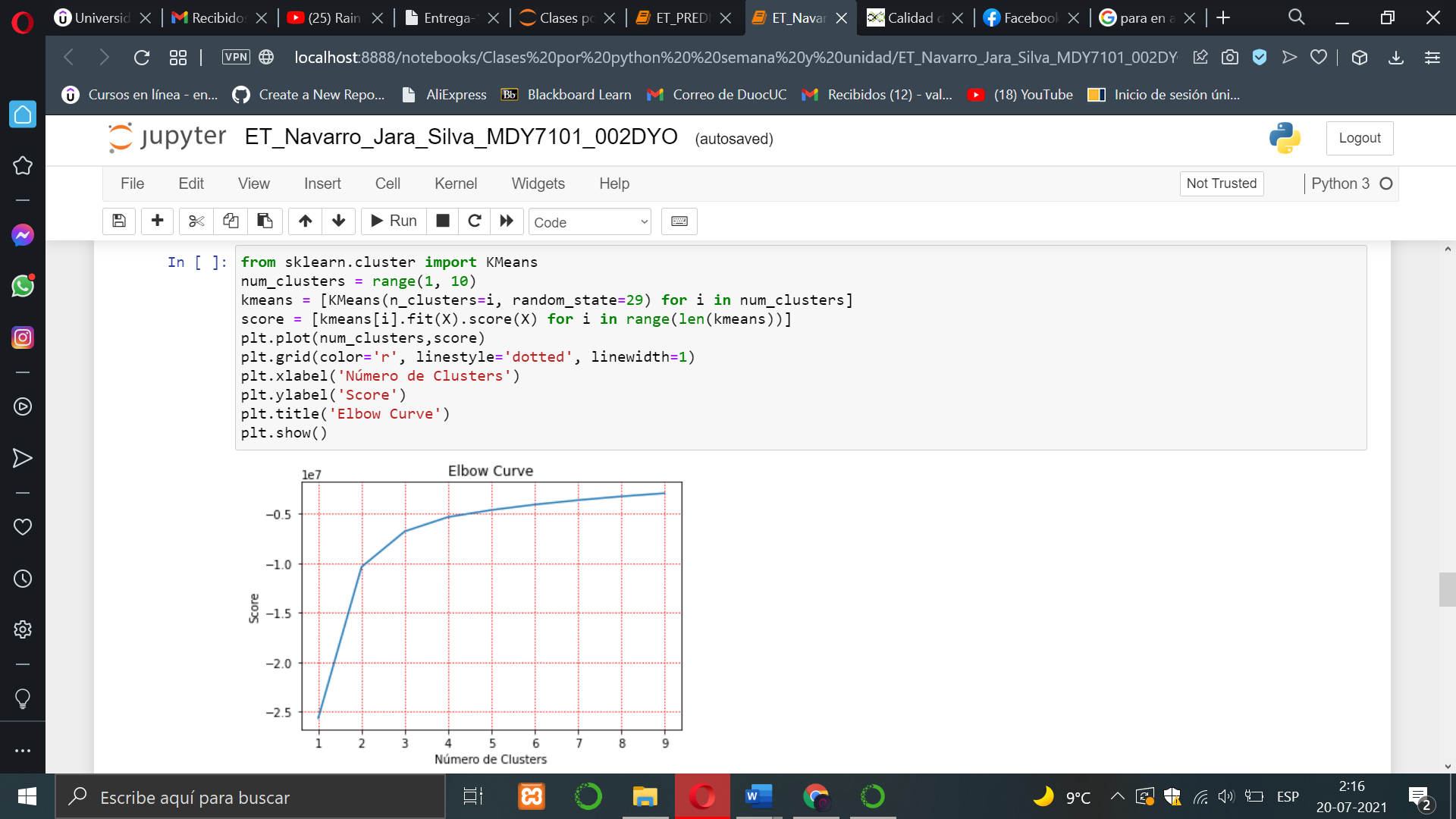


**Segmentación**

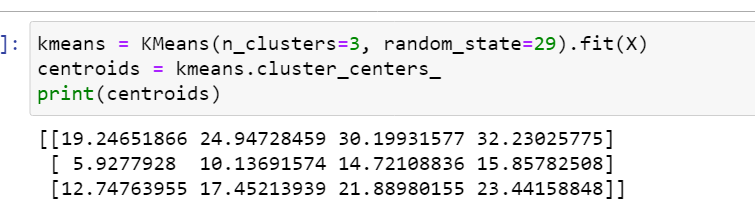
**Datos no normalizados**

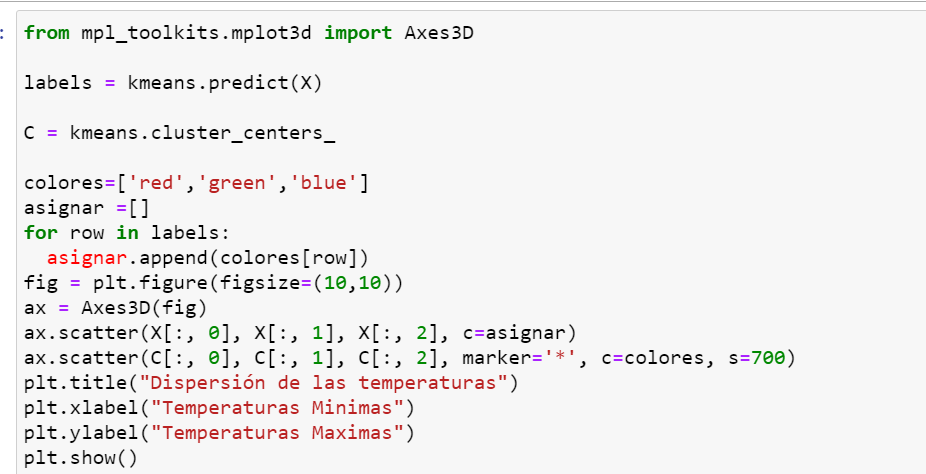


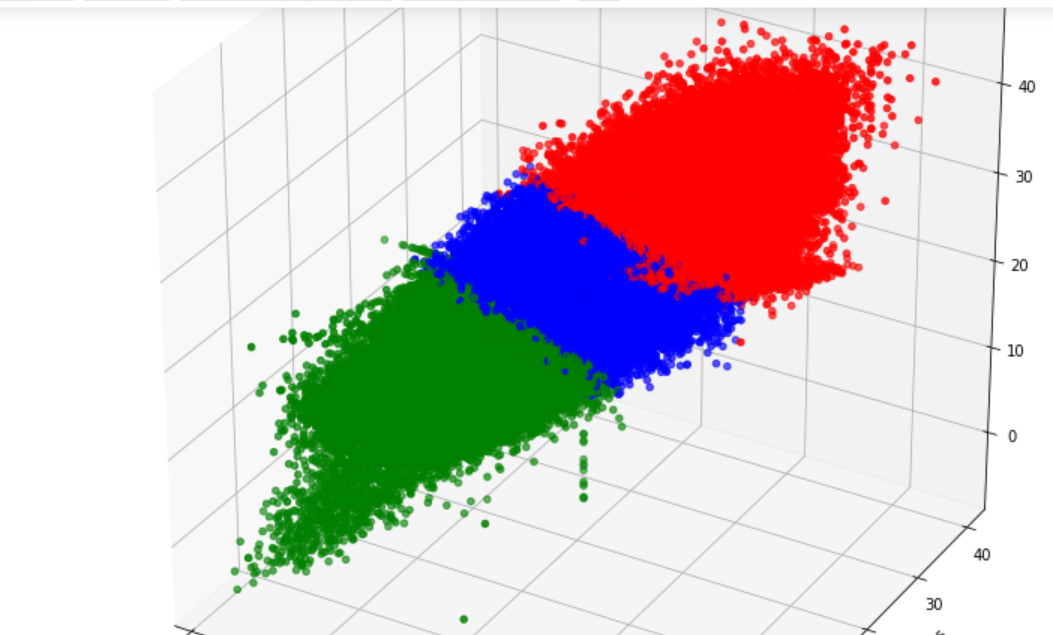
Se asigna el valor de K, en este caso decidimos usar las temperaturas en general, para así obtener una mejor segmentación



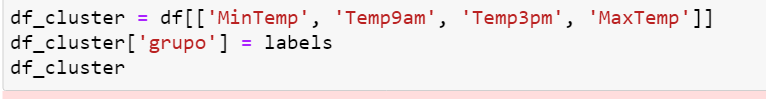
Se usará el cluster 3 ya que es donde la línea se normaliza

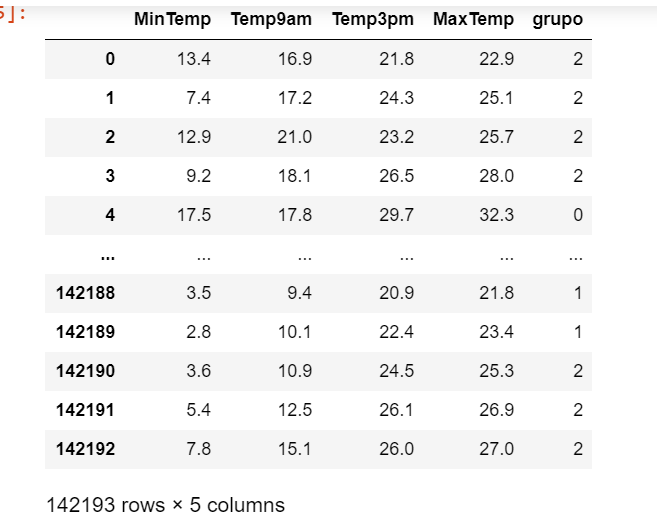


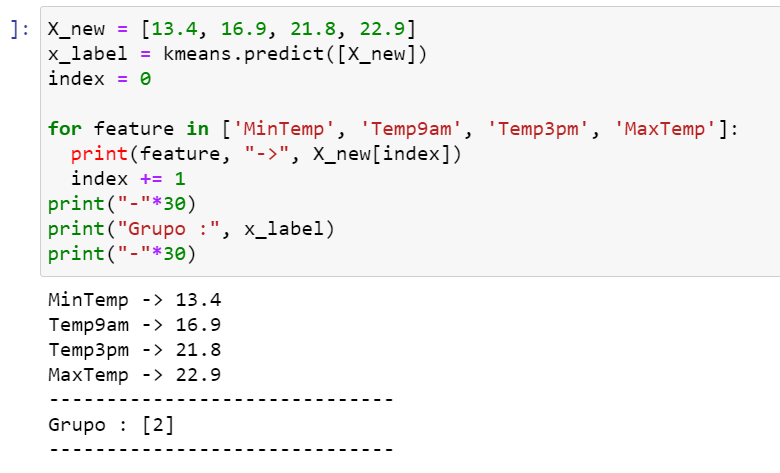




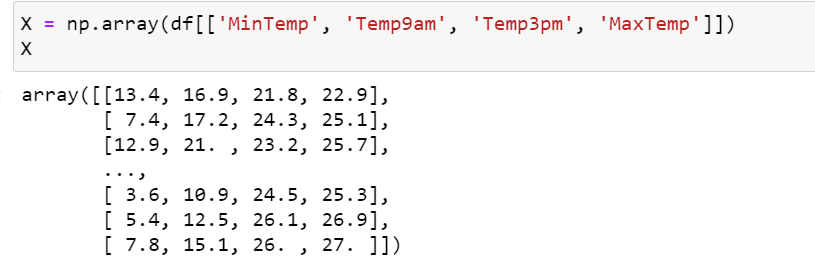
En este gráfico se muestra la dispersión de las temperaturas dentro del cluster N°3



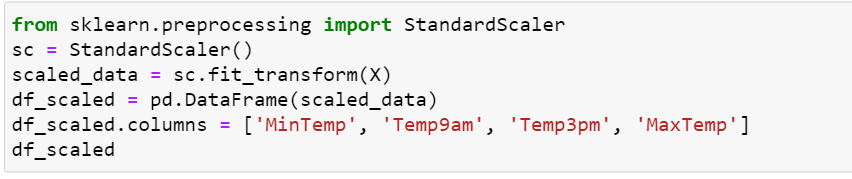


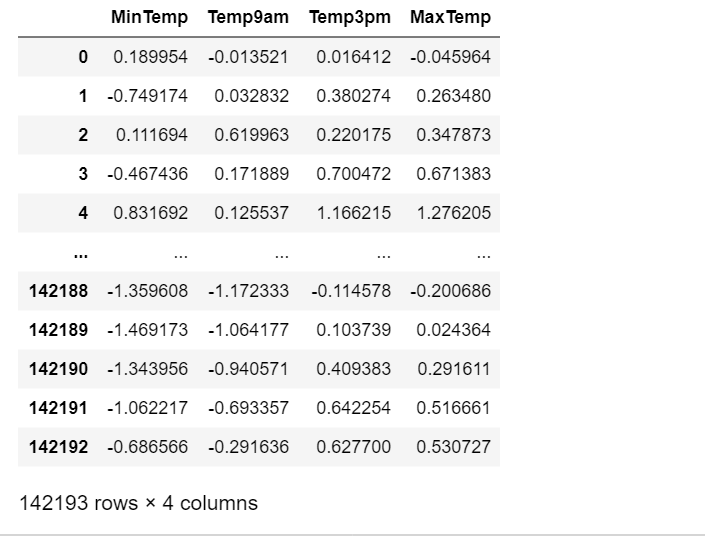


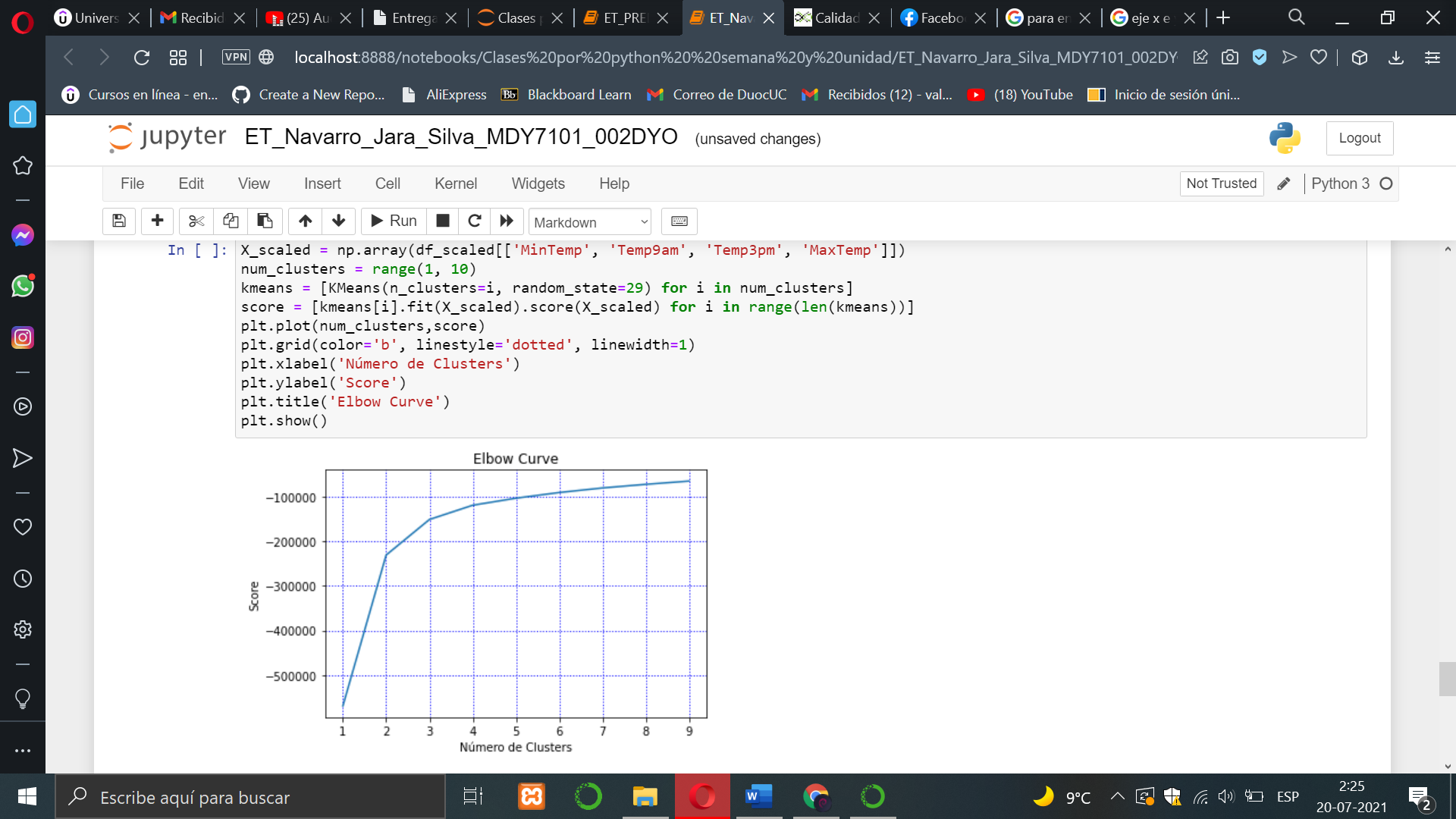
### **Datos Normalizados**



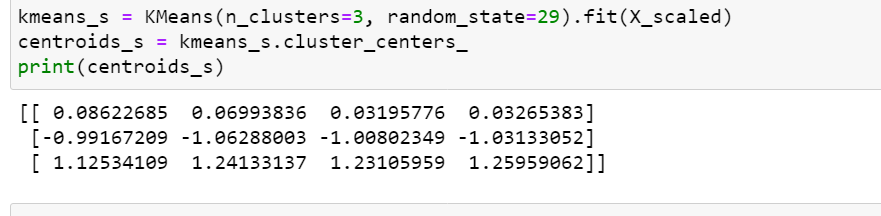
Se normalizan los datos a utilizar

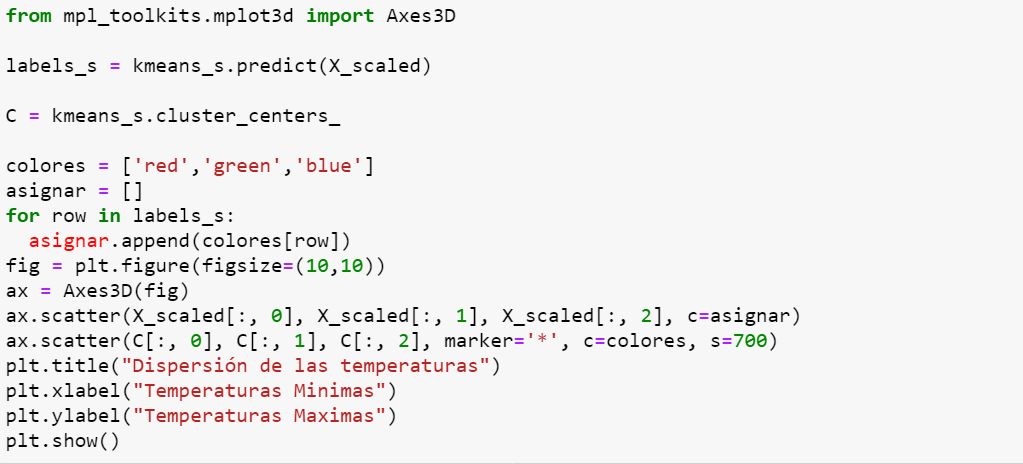


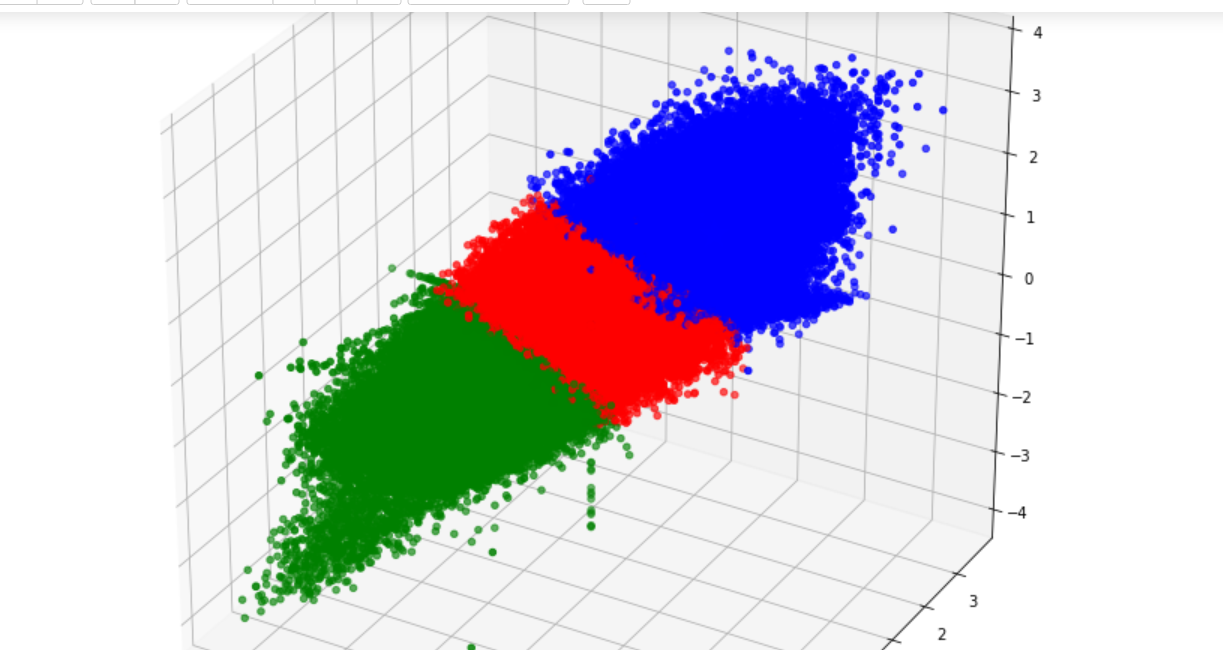




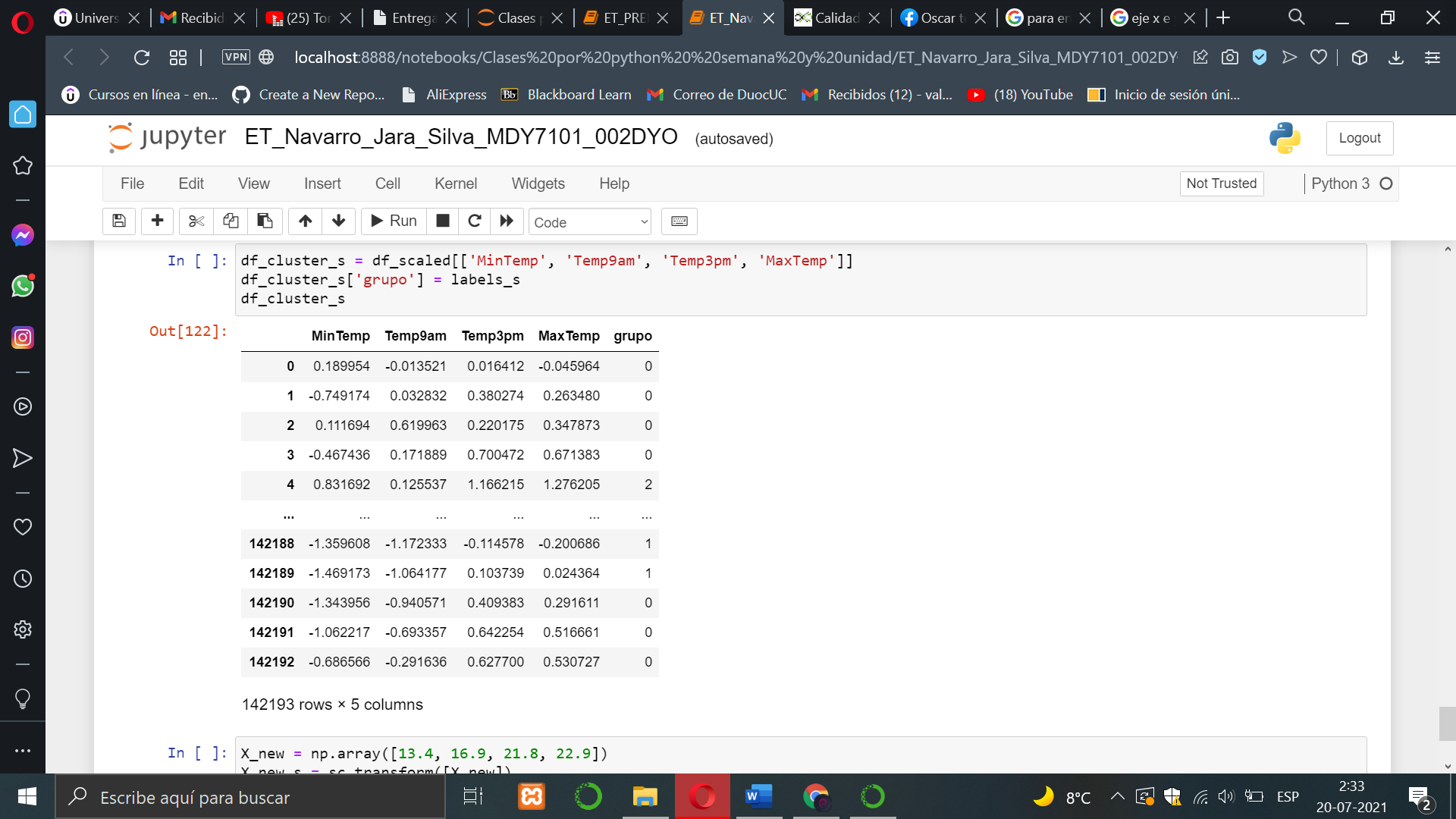
Se usará el cluster 3 ya que es donde la línea se normaliza

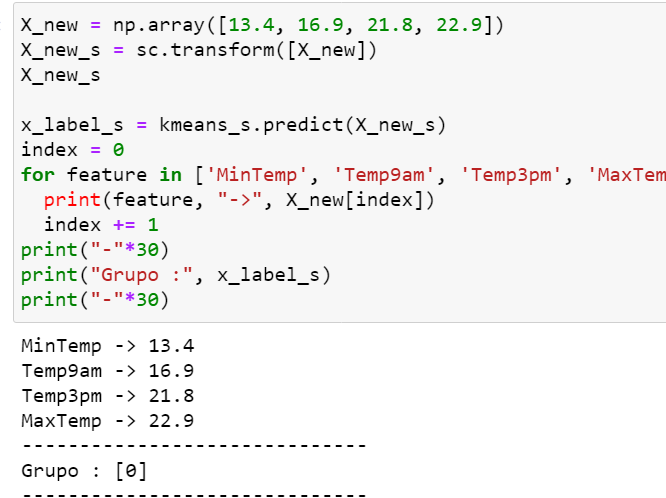






En este gráfico se muestra la dispersión de las temperaturas dentro del cluster N°3





# 

# **4- Razonamiento y justificación respecto al análisis de datos**

Los datos utilizados dentro del proyecto tienen la funcionalidad suficiente para poder generar información importante, como podría ser la velocidad de las rafagas de viento (WindGustSpeed) que se le podría dar la utilidad en el negocio de las energías renovables encontrando el mejor lugar donde se pueda crear una granja eólica.

Pero no solamente se le dio un enfoque centrado en el negocio ya que igualmente se busca ayudar a la gente, intentando predecir qué día lloverá por ejemplo o en qué momento del día hay mayores velocidades en el viento para tomar precauciones.

Como bien se puede ver en el proyecto se podría decir que se hizo con la intención de sacar el mayor provecho a los datos entregados intentando siempre mantener un enfoque en ayudar al cliente.

# **5- Definición de fases, tareas y entregables a partir del conjunto de datos y proyecto en general**

**Fases y tareas:**

**1-Exploracion y analisis de datos**

* Importación de librerías necesarias para las siguientes tareas.
* Exploración de los datos.

**2-procesamiento de datos**

* Limpieza de datos nulos.
* Transformación de datos.

**3-visualizacion y analisis de comportamiento de los datos**

* Aplicación de gráficos útiles para comprender el comportamiento de los datos.
* Estudio de comparación, tendencias, partes de un total, distribución y correlación de datos,

**4-Machine learning**

* Predicción
* Clasificación
* Segmentación

**Entregables:**

1-Notebook con el código Python.

2-Informe de tipo ejecutivo.

# **6- Sugerir decisiones estratégicas para apoyar la organización**

Analizando el comportamiento de los datos viendo sus gráficas, se podría sugerir a la organizacion utilizar los datos de las rafagas de viento para la creacion de granjas eólicas en zonas como Penrith que tiene las velocidades de rafagas más altas en toda Australia, con una orientación a las direcciones NorOeste y Oeste NorOeste.

Aparte, con la clasificación se puede hacer una “predicción” para saber si lloverá el próximo día, usando esta información en conjunto con los gráficos, los cuales indican que en los días lluviosos hay menos rafagas de viento, ésto se podría usar en conjunto para tener una mayor efectividad en la energía eólica, quizás manteniendo un margen en los días lluviosos para impedir daño en la infraestructura y ocupar toda la potencia en los días soleado para una mayor eficiencia.