Análisis del impacto ambiental de la Industria Ganadera

Profesor:

Mauricio Araya

Integrantes:

Mario Araya F. 201630003-1
Marcelo Villablanca 201530009-7
Simón Rivera 201130030-0

09/08/2020

INTRODUCCIÓN

> “En lo que respecta al cambio climático y al calentamiento global, lo que ignoramos supera con mucho lo que sabemos. Para empeorar aún más las cosas, ni siquiera sabemos si la información de la que disponemos es realmente fiable. Tampoco comprendemos de forma cabal cuáles son las soluciones que tienen un carácter razonable".(Nathan Sykes en OpenMind BBVA, Big Data y la lucha contra el cambio climático)

El calentamiento global es un problema que hoy en día les concierne a todas las personas porque aunque no estén concientes han sido parte del problema todo este tiempo, consumiendo energía eléctrica generada a partir de combustibles fósiles, productos industriales y alimentos que generan decenas de toneladas de gases de efecto invernadero(GEI) al año.

Es sabido que hay muchos sistemas que interactúan en el calentamiento global y muchos de estos sistemas dependen unos de otros para desarrollarse, pero hay uno en particular que forma parte del día a día de la gran mayoría de culturas humanas en el planeta, hablamos del consumo de carne y la industria ganadera. Es por esto que en el siguiente informe se procederá a analizar la participación de esta industria en el calentamiento global y como punto inicial se realizará la siguiente pregunta, ¿Cómo contribuye la industria ganadera al calentamiento global?

INVESTIGACIÓN

La ganadería no es la única industria que genera gases de efecto invernadero, por esta razón se procederá a realizar una investigación a fondo sobre el problema, donde se abordarán los entes más destacados en este sistema de contaminación y negligencia hacia el medio ambiente.

Por un lado, el 14,5% de las emisiones globales de GEI antropogénicas proviene de la ganadería y el ganado vacuno (carne y leche) es responsable de aproximadamente dos tercios de esta cifra.

La deforestación y degradación de los bosques, causada por la expansión agrícola, la conversión de los bosques en pastizales, la tala destructiva y los incendios representan el 11% de las emisiones de GEI.

El sector ganadero consume anualmente 6,000 millones de toneladas de alimentos entre forrajes, granos, piensos y otros materiales, incluyendo un tercio de la producción mundial de cereales. El 86% de la ingesta animal se compone de materiales que no son de consumo humano, además las especies monogástricas representan un 72% del consumo mundial de cereales del sector ganadero, mientras que los forrajes y la vegetación constituyen más del 57% de la ingesta total de las especies de rumiantes. En consecuencia de lo anterior el sector ganadero contribuye significativamente al total de emisiones humanas de GEI. Se estima que las cadenas de producción ganadera emitieron globalmente un total de 8,1 gigatoneladas de CO_2-eq(Dióxido de carbono equivalente) en 2010 (usando los últimos índices de potencial de calentamiento del IPCC(The Intergovernmental Panel on Climate Change): 298 para N_2O y 34 para CH_4). El metano (CH_4) representa un 50% del total. El óxido nitroso (N_2O) y el dióxido de carbono (CO_2) muestran porcentajes similares, siendo éstos un 24 y un 26 por ciento, respectivamente.

La fuente más importante de **metano**(CH4) es la **descomposición de materia orgánica** en sistemas biológicos, siendo en las actividades agrícolas relacionadas con la fermentación entérica como consecuencia del proceso digestivo de los herbívoros, la descomposición en condiciones anaerobias (sin oxígeno) del estiércol generado por especies pecuarias, los cultivos de arroz bajo riego y las quemas de sabanas y residuos agrícolas. Por otro lado, en lo que respecta a la disposición de residuos sólidos sería, el tratamiento anaerobio de aguas residuales domésticas e industriales. Otra fuente importante de metano está relacionada con la producción y distribución de gas natural y petróleo y en la explotación de carbón mineral. El efecto de las emisiones de metano por fermentación intestinal de los rumiantes es bastante grande a nivel global y se estima que esta fuente produce hasta el 37% del metano presente en la atmósfera

El **óxido nitroso**(N2O), cuyas fuentes son de carácter natural y antropogénico, contribuye con

cerca del 6% del forzamiento del efecto invernadero. Sus fuentes incluyen los océanos, la quema de combustibles fósiles, biomasa y la agricultura. El óxido nitroso es inerte en la troposfera. Su principal sumidero es a través de las reacciones fotoquímicas en la estratosfera que afectan la abundancia de ozono estratosférico. **La fuente más importante de óxido nitroso son las emisiones generadas por suelos agrícolas y en menor grado por el consumo de combustibles fósiles para generar energía y las emitidas por descomposición de proteínas de aguas residuales domésticas**. Las emisiones de óxido nitroso generadas por los suelos agrícolas se deben principalmente al proceso microbiológico de la nitrificación y desnitrificación del suelo. Se pueden distinguir tres tipos de emisiones: las directas desde el suelo, las directas de óxido nitroso del suelo debido a la producción animal (pastoreo) y las indirectas generadas por el uso de fertilizantes.

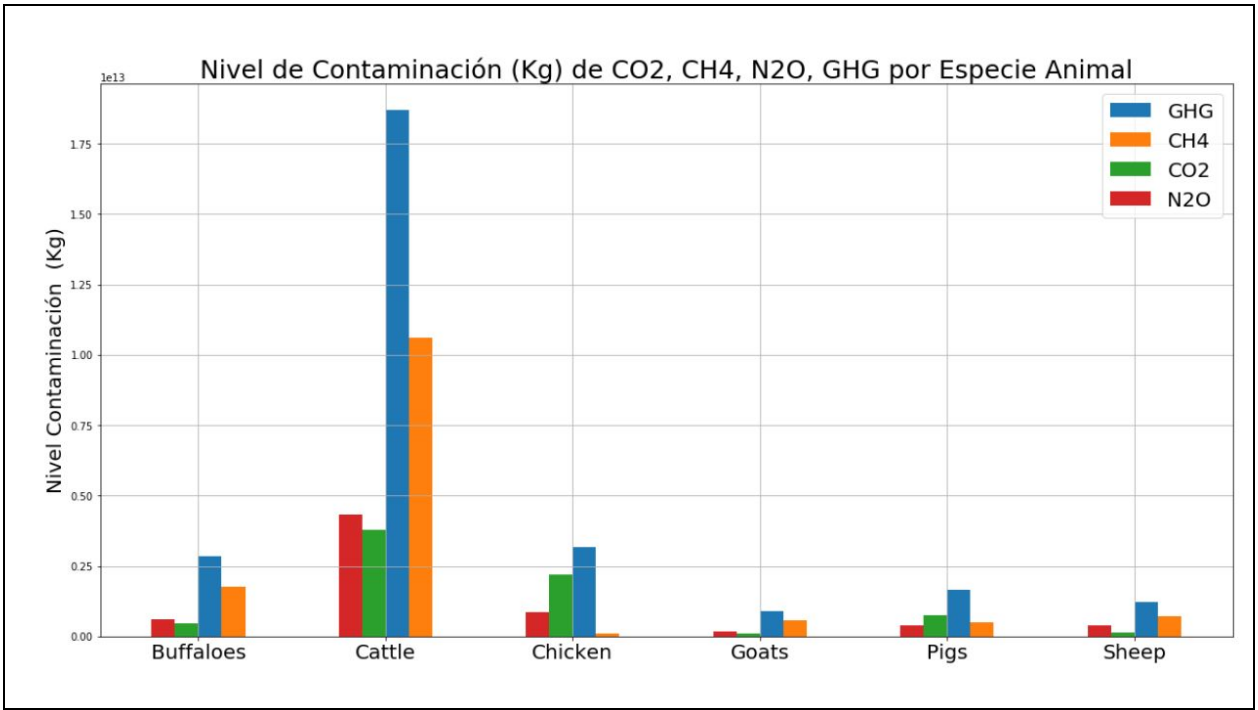
El **dióxido de carbono**(CO_2) es uno de los gases traza más comunes e importantes en el sistema atmósfera-océano-Tierra, es el más importante GEI asociado a actividades humanas y el segundo gas más importante en el calentamiento global después del vapor de agua. Este gas tiene fuentes antropogénicas y naturales. Dentro del ciclo natural del carbono, el CO_2 juega un rol principal en un gran número de procesos biológicos. **En relación a las actividades humanas el CO_2 se emite principalmente, por el consumo de combustibles fósiles** (carbón, petróleo y sus derivados y gas natural) **leña para generar energía, por la tala y quema de bosques** (según la FAO, el 26% de la superficie terrestre se destina al pastoreo, y la producción de forrajes requiere de cerca de una tercera parte del total de la superficie agrícola; La principal causa de deforestación en América Latina se debe, justamente, a la expansión de tierras para el pastoreo; El 70% de los bosques amazónicos se usan como pastizales) **y por algunos procesos industriales como la fabricación del cemento.**

La deforestación y degradación causadas por el ser humano, también hacen que los bosques sean más propensos a los incendios porque crean condiciones más secas en el clima. Por otro lado, la región de la Amazonía Brasileña, perdió 3,7 millones de hectáreas (9,1 millones de acres) de cobertura arbórea durante el año civil de 2016, casi tres veces más de lo que había perdido en 2015. **Desde 2005 a 2015 una de cada diez hectáreas perdidas del Amazonas han sido a causa de la minería**. La minería ilegal está causando efectos devastadores en la Amazonía por la presencia de dragas, barcas y otros equipos utilizados para la extracción de oro que acaban con los bosques, así como por el uso indiscriminado de mercurio que genera daños a la salud de las poblaciones locales (principalmente indígenas) y afecta a ríos y peces. Según la Red Amazónica de Información Socio Ambiental y Georreferenciada (RAISG), hay más de 450 minas ilegales en la Amazonía brasileña. Además, según una investigación de la ONG internacional Global Witness, en 2018 fueron asesinados 20 defensores de la tierra y del medio ambiente en Brasil. A

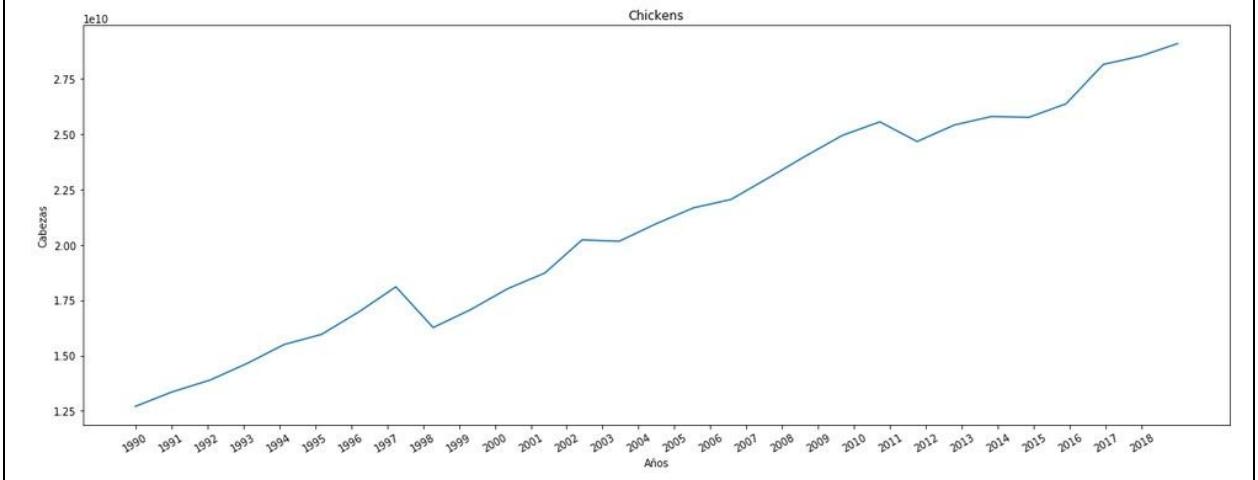
escala mundial, la organización sin fines de lucro citó la minería como el sector más mortífero, con 46 asesinatos cuantificados ese mismo año.

Y por último pero no menos importante es el impacto que tienen las industrias encargadas de la generación de energía eléctrica. **El 76,3 % de la indispensable energía eléctrica se obtiene de fuentes no renovables, y el restante 23,7 %, de energías verdes** (de entre las cuales la hidráulica representó el 16,6 %).

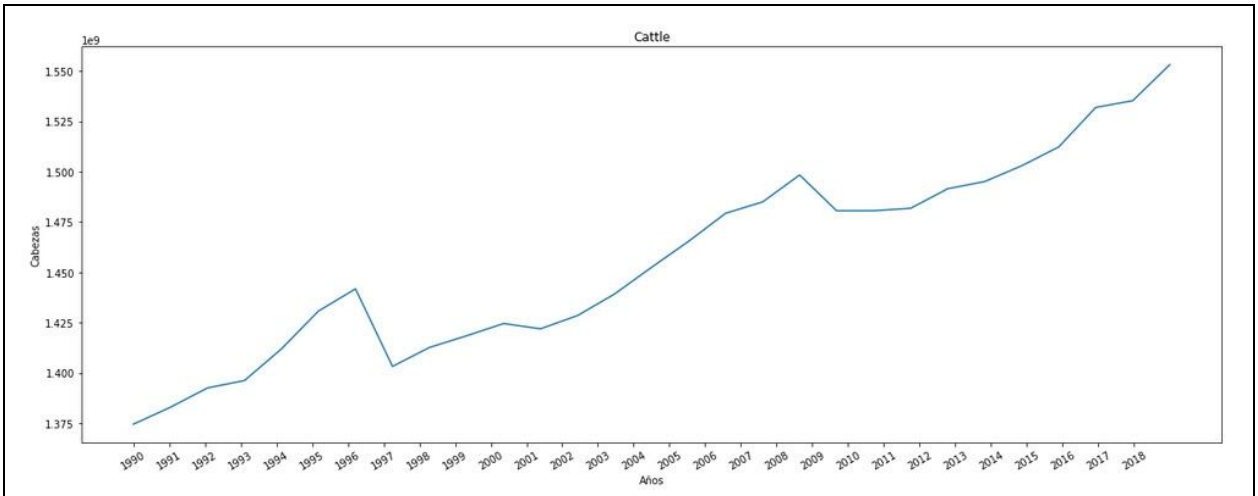
Teniendo todo lo anterior en cuenta, el abuso de la quema de combustibles fósiles, la descomposición de materia orgánica, la necesidad de nuevos espacios para la agricultura y la explotación de recursos naturales está directamente relacionada al aumento de la población en el planeta, para satisfacer las "necesidades" de los humanos modernos y su estilo de vida.



Nivel de contaminación en Kg de GEI de la industria ganadera por animal, según GLEAM 2010.



Cantidad de pollos/gallinas destinados a la industria desde 1990-2018.



Cantidad de vacas destinadas a la industria desde 1990-2018.

HIPÓTESIS

La industria ganadera es una de las principales responsables del aumento de gases de efecto invernadero(GEI) en nuestro planeta.

MATERIAL

1. FAOSTAT Página oficial de la FAO(Organización de las Naciones Unidas para la Alimentación y la Agricultura)
2. http://www.fao.org/faostat/es/#data/RL (Agricultura)
3. http://www.fao.org/faostat/es/#data/QA (Ganadería)
4. http://www.fao.org/faostat/en/#data/GF (Área forestal)
5. Consumo de energía eléctrica (kWh per cápita) | Data
6. Uso de energía (kg de equivalente de petróleo per cápita) | Data
7. World population by region
8. Área selvática (kilómetros cuadrados) | Data
9. Gold Price Historical Data | Gold Price History
10. Per capita CO_2 emissions
11. Forest Fires in Brazil
12. Methane emissions by sector
13. Nitrous oxide emissions
14. Average global mean temperature anomalies in degrees Celsius (GCAG base period: 20th century average.)

DATOS

Características	Descripción
Tierra agrícola	Típicamente tierra dedicada a la agricultura, el uso sistemático y controlado de otras formas de vida (particularmente la cría de ganado y la producción de cultivos) para producir alimentos para los humanos en Kha.
Tierra de cultivo	Se refiere a los cultivos que requieren re-plantación anual o barbecho o pastos utilizados para tales cultivos dentro de un período de cinco años en Kha.
Tierras arables	Tierra que puede ser usada para la agricultura en Kha.
Tierras dedicadas a praderas y pastizales permanentes	En Kha.
Tierras destinadas a cultivos permanentes	En Kha.
Tierras en barbecho	Tierras en periodo de descanso para recuperar y almacenar materia orgánica y humedad, en Kha.

Tierras destinadas a cultivos temporales	En Kha.
Tierras destinadas a praderas y pastizales temporales	En Kha.
Praderas y pastizales permanentes naturales	En Kha.
Praderas y pastizales permanentes cultivados	En Kha.
Tierras agrícolas bajo cubiertas protectoras	Tierras protegidas bajo estructuras que proveen condiciones óptimas al cultivo. Ej: Invernaderos, malla sombra, macrotúnel, etc. En Kha.
Chickens	Número de cabezas de gallinas/pollos destinadas a la industria ganadera. (Avicultura).
Cattle	Número de vacas en cabezas destinadas a la industria.
Goats	Número de cabras en cabezas destinadas a la industria.
Sheep	Número de ovejas en cabezas destinadas a la industria.
Buffaloes	Número de búfalos en cabezas destinadas a la industria.
Pigs	Número de cerdos en cabezas destinadas a la industria.
Uso de energía de combustibles fósiles por kg equivalente de petróleo por cabeza	El uso de energía se refiere al consumo de energía primaria antes de la transformación en otros combustibles finales, lo que equivale a la producción nacional más las importaciones y las variaciones de existencias, menos las exportaciones y los combustibles suministrados a barcos y aviones afectados al transporte internacional. En Kg eq de petróleo per cápita.
Uso de energía eléctrica por kilo Watts por hora por cabeza	En KWh per cápita.
N incendios(Amazonía BR)	Número de incendios ocurridos en la Amazonía perteneciente a Brasil (64.4% del área total).
Gold price	Precio del oro. Característica introduce la participación de las mineras ilegales en la Amazonía.

CH4 in Agriculture	Gas metano producido por la agricultura en toneladas de CO2 equivalente[tCO2eq].
CH4 in Fugitive emissions	Las emisiones fugitivas son emisiones de gases no intencionadas o irregulares, principalmente de maquinaria o en este caso flatulencias en tCO2eq.
CH4 in Waste	Basura o estiércol que se descompone en tCO2eq.
CH4 in Land use change and forestry	Cambio de tierra y silvicultura absorción en tCO2eq.
CH4 in Industry	Toneladas de CO2eq emitidas por industrias.
CH4 in Other fuel combustion	Toneladas de CO2eq emitidas por otro uso de combustibles fósiles.
Population	Población humana.
Forest area	Área forestal en Kha.
Área Selvática	En Kha.
NO2 Total including land use change and forestry	Óxido nitroso total emitido en tCO2eq, incluyendo el absorbido por el cambio de tierra y silvicultura.
CO2 Total	En toneladas de emisión.
CH4 Total	En toneladas de emisión, incluyendo el absorbido por el cambio de tierra y silvicultura.

PROCEDIMIENTO

1. Se pre-procesan los 13 Datasets para obtener uno con todos los datos a analizar.
2. Analizar el dataset a partir de la descripción de pandas y desplegar su información, es decir, forma de la matriz, tipos de datos y lista de características.
3. Pequeño reajuste en el Dataset. Se agrega columna **CH4 Total**, sumando todas las columnas de emisión de **CH4** por sector.
4. Se procede a estandarizar los datos y se imprime la descripción de pandas y su información.
5. Se obtiene matriz de correlaciones.

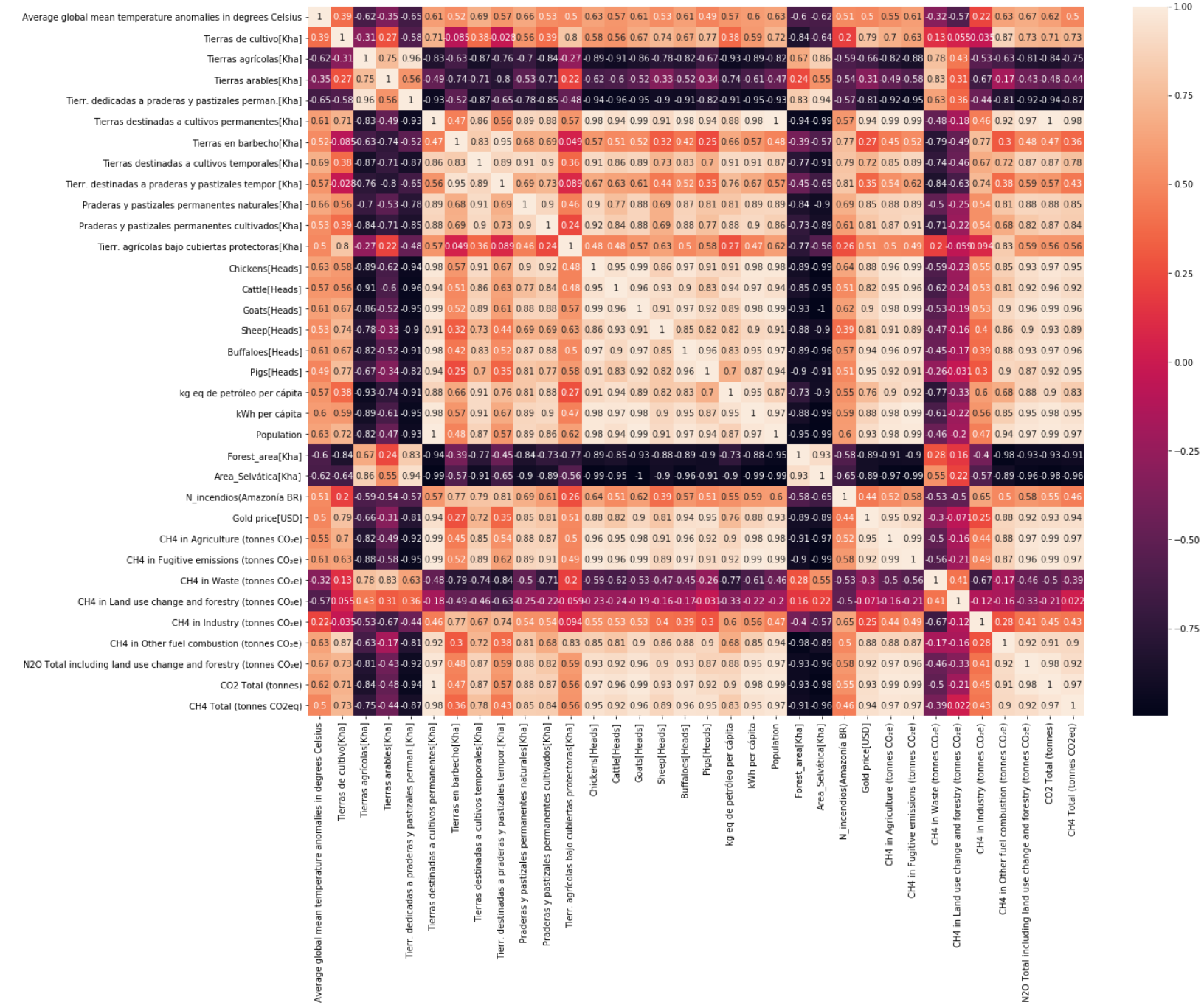
6. Se generan gráficos de todas las características versus **CH4**, **N2O** y **CO2**. Todos estos mostrando el comportamiento de la característica “population” a través de colores con **Seaborn pairplot**.
7. Se aplica análisis de **componentes principales(PCA)**. Para esto se extraen las características de interés, **GEI** del Dataset. Luego se procede a aplicar la reducción de dimensionalidad de **PCA** a 5 características y graficar la proporción de la varianza por cada componente principal obtenido del entrenamiento(ejecución) de **PCA**, junto con la matriz de correlación de las componentes y las características del Dataset para analizar su relación, además verificar la perpendicularidad entre las componentes(correlación 0) con el fin de comprobar el funcionamiento adecuado de **PCA**.
8. Al ir todo bien, se grafican las 5 componentes mostrando su relación con los **GEI**. Para esto se utiliza **Seaborn pairplot** para cada gas en específico.
9. Luego al tener las componentes principales listas, se aplica un algoritmo de clustering para identificar posibles grupos que participen de forma más activa en cada **GEI**.
10. Este algoritmo será **K-means**, el cual se ejecutará en 2 ocasiones con diferentes variaciones del algoritmo de esperanza-maximización (**EM**), “full” y “elkan”.
11. Se grafican los resultados de **K-means** con cien iteraciones y ambas variaciones del algoritmo **EM**.
12. Se procede a desplegar los datos mostrando cada uno su cluster correspondiente.
13. Se aplica otro algoritmo de clustering, **Agglomerative Clustering** y se grafican los resultados.
14. Se inicia el procedimiento para estimar un modelo apropiado a través de **regresión**.
15. Se seleccionan variables target y se dejan aparte del dataset.
16. Para cada una de las variables target se visualiza su correlación y se dejan las características con correlación absoluta mayor a 0.5.
17. Se inicia la prueba de modelos, utilizando **Ridge**, donde se evalúa para distintas constantes de regularización.
18. Se gráfica el error cuadrático medio para los distintos parámetros de regularización, la precisión para distintos parámetros de regularización y el peso que el modelo de dio a cada característica.
19. Se realiza lo anterior descrito para el modelo **Lasso**.
20. Se prueba con modelo **Random Forest**, por lo que se deben estimar sus parámetros.
21. Para realizar la estimación de los mejores parámetros se utiliza **Random Search**.
22. Se realiza comparación con modelo **Random Forest** por defecto.
23. Se grafica la contribución de cada característica.

ANÁLISIS Y RESULTADOS

Lista de características del Dataset.

```
Int64Index: 17 entries, 1998 to 2014
Data columns (total 34 columns):
 #   Column                                                    Non-Null Count  Dtype
---  ------                                                    --------------  -----
 0   Average global mean temperature anomalies in degrees Celsius  17 non-null     float64
 1   Tierras de cultivo[Kha]                                   17 non-null     float64
 2   Tierras agrícolas[Kha]                                    17 non-null     float64
 3   Tierras arables[Kha]                                      17 non-null     float64
 4   Tierr. dedicadas a praderas y pastizales perman.[Kha]     17 non-null     float64
 5   Tierras destinadas a cultivos permanentes[Kha]            17 non-null     float64
 6   Tierras en barbecho[Kha]                                  17 non-null     float64
 7   Tierras destinadas a cultivos temporales[Kha]             17 non-null     float64
 8   Tierr. destinadas a praderas y pastizales tempor.[Kha]    17 non-null     float64
 9   Praderas y pastizales permanentes naturales[Kha]          17 non-null     float64
 10  Praderas y pastizales permanentes cultivados[Kha]         17 non-null     float64
 11  Tierr. agrícolas bajo cubiertas protectoras[Kha]          17 non-null     float64
 12  Chickens[Heads]                                           17 non-null     float64
 13  Cattle[Heads]                                             17 non-null     float64
 14  Goats[Heads]                                              17 non-null     float64
 15  Sheep[Heads]                                              17 non-null     float64
 16  Buffaloes[Heads]                                          17 non-null     float64
 17  Pigs[Heads]                                               17 non-null     float64
 18  kg eq de petróleo per cápita                              17 non-null     float64
 19  kWh per cápita                                            17 non-null     float64
 20  Population                                                17 non-null     float64
 21  Forest_area[Kha]                                          17 non-null     float64
 22  Area_Selvática[Kha]                                       17 non-null     float64
 23  N_incendios(Amazonia BR)                                  17 non-null     float64
 24  Gold price[USD]                                           17 non-null     float64
 25  CH4 in Agriculture (tonnes CO2e)                          17 non-null     float64
 26  CH4 in Fugitive emissions (tonnes CO2e)                   17 non-null     float64
 27  CH4 in Waste (tonnes CO2e)                                17 non-null     float64
 28  CH4 in Land use change and forestry (tonnes CO2e)         17 non-null     float64
 29  CH4 in Industry (tonnes CO2e)                             17 non-null     float64
 30  CH4 in Other fuel combustion (tonnes CO2e)                17 non-null     float64
 31  N2O Total including land use change and forestry (tonnes CO2e)  17 non-null     float64
 32  CO2 Total (tonnes)                                        17 non-null     float64
 33  CH4 Total (tonnes CO2eq)                                  17 non-null     float64
dtypes: float64(34)
```

Luego de estandarizar los datos se procede a generar la matriz de correlaciones.



Observando la matriz de correlaciones se podrá notar una ligera inclinación por parte del **óxido nitroso**(N2O) por la quema de combustibles fósiles con una correlación de 0.88 con los Kg de petróleo per cápita que supera a la del **CH4** 0.85, sin mencionar al **CO2** es tiene una correlación un poco mayor 0.9. Además se puede notar una alta correlación con las tierras destinadas a

cultivos permanentes al igual que el resto de los gases.

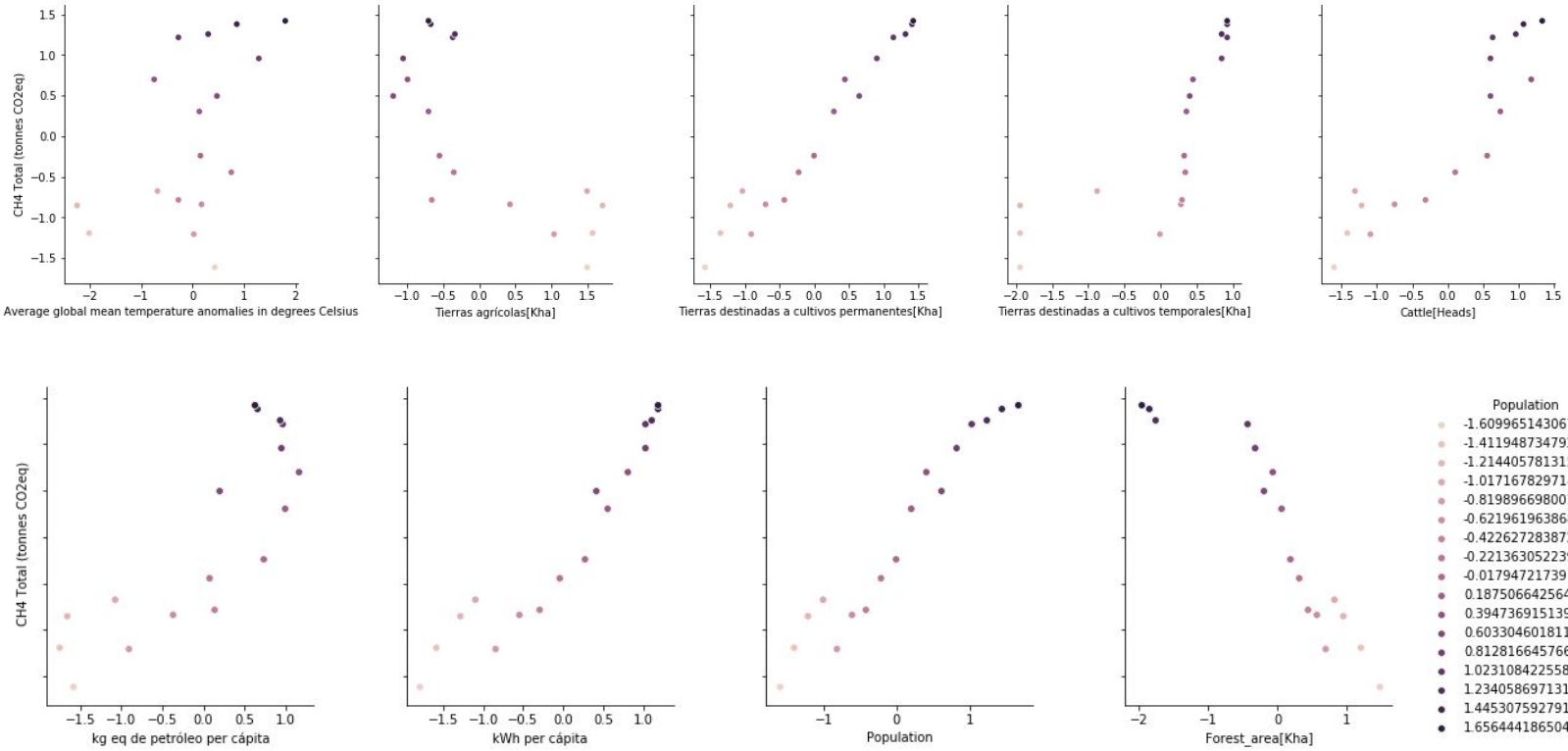
También se puede notar que el **N2O** junto con el **CO2** tienen las correlaciones negativas más "altas", reafirmando su contribución por combustión de combustibles fósiles y no fósiles.

Por otro lado, con respecto al **gas metano**(CH4), se puede notar una leve inclinación hacia la ganadería teniendo correlaciones ligeramente mayores al **NO2** en este sector. También cabe destacar que aunque sea mínimo, presenta una correlación mayor al **NO2** en lo que respecta a agricultura en cultivos permanentes.

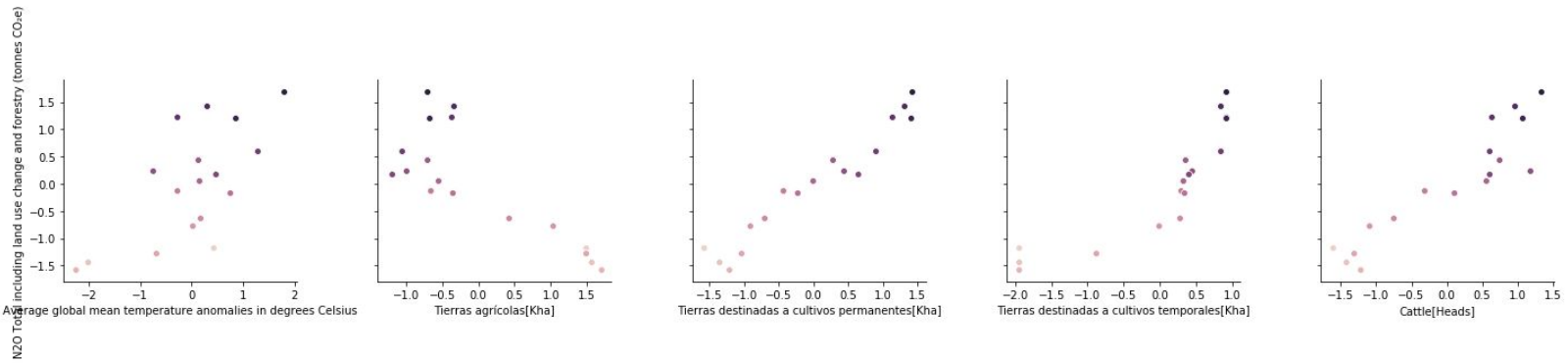
Por último tenemos al **dióxido de carbono**(CO2) que es un gas que está presente en todos los procesos y con gran correlación. En la agricultura tenemos una correlación de 1 con los cultivos permanentes, en la ganadería incluso supera al **CH4** y en la quema de combustibles supera al **N2O**, teniendo además una correlación de 0.99(la más alta) con la población humana global.

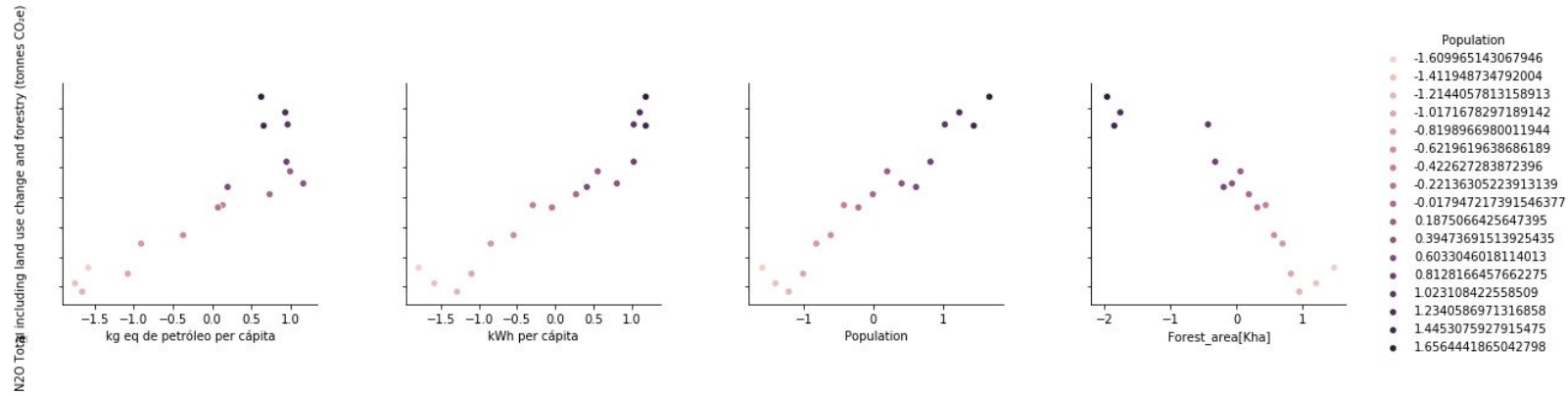
A continuación los gráficos representan las entidades más importantes del Dataset versus los tres gases de efecto invernadero.

- **<u>Metano(CH4):</u>**

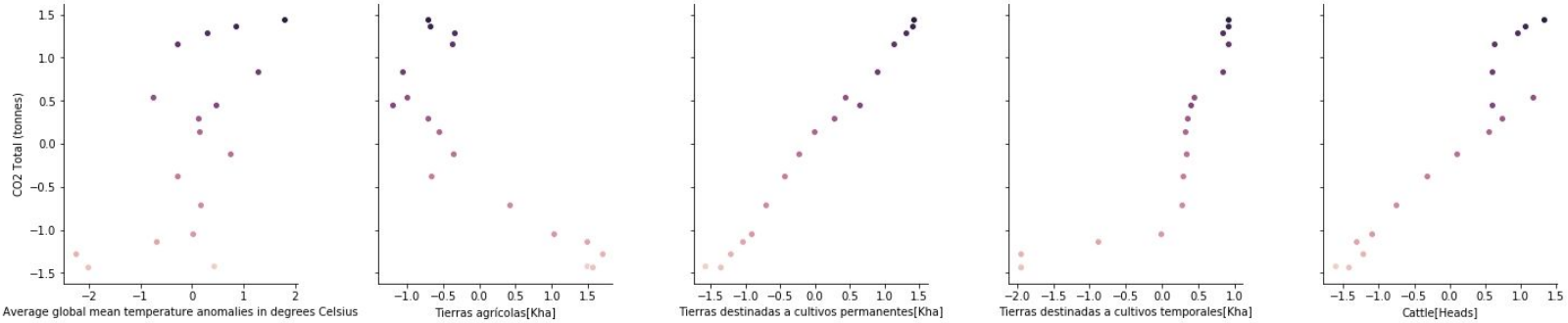


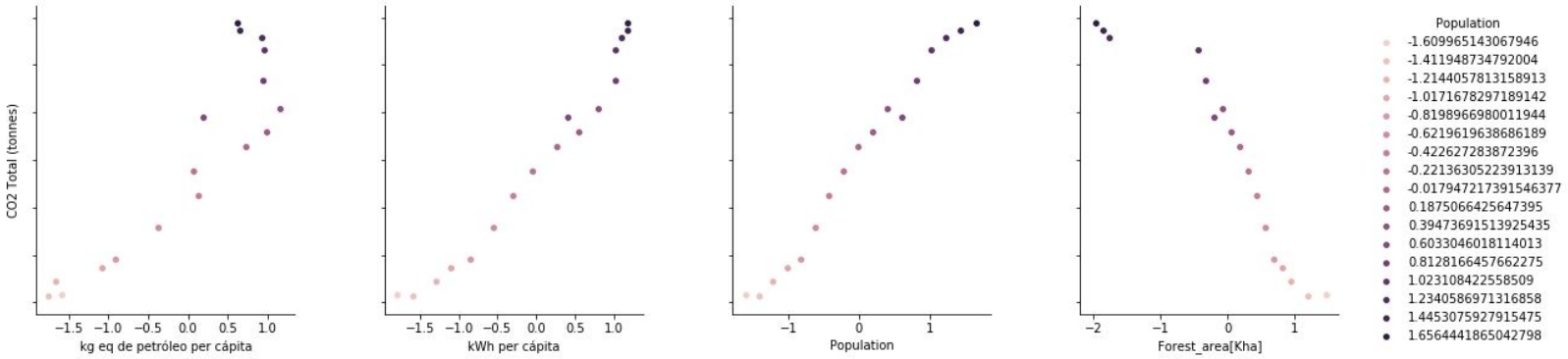
- **Óxido nitroso(N2O):**





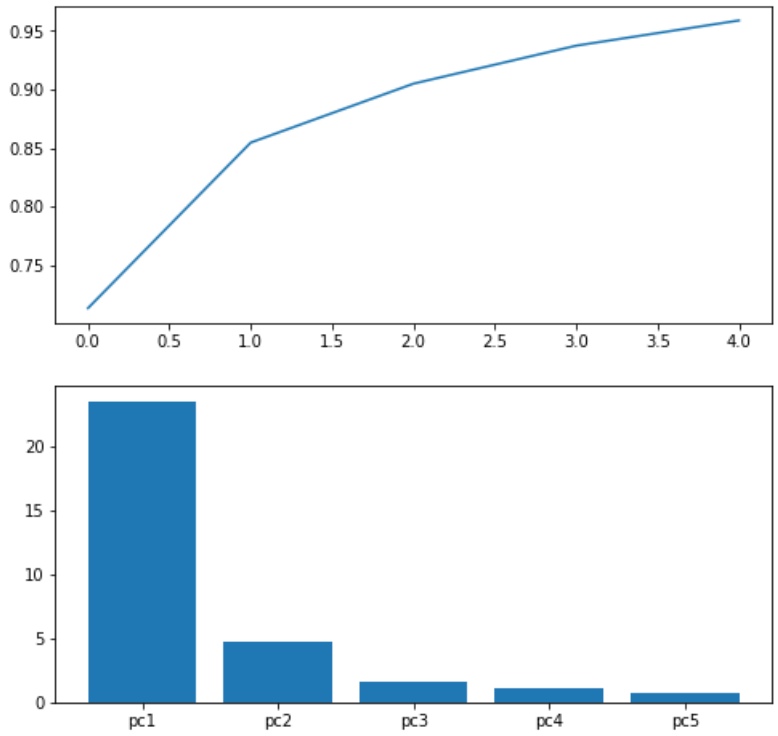
- **Dióxido de carbono:**





PCA(Principal component analysis):

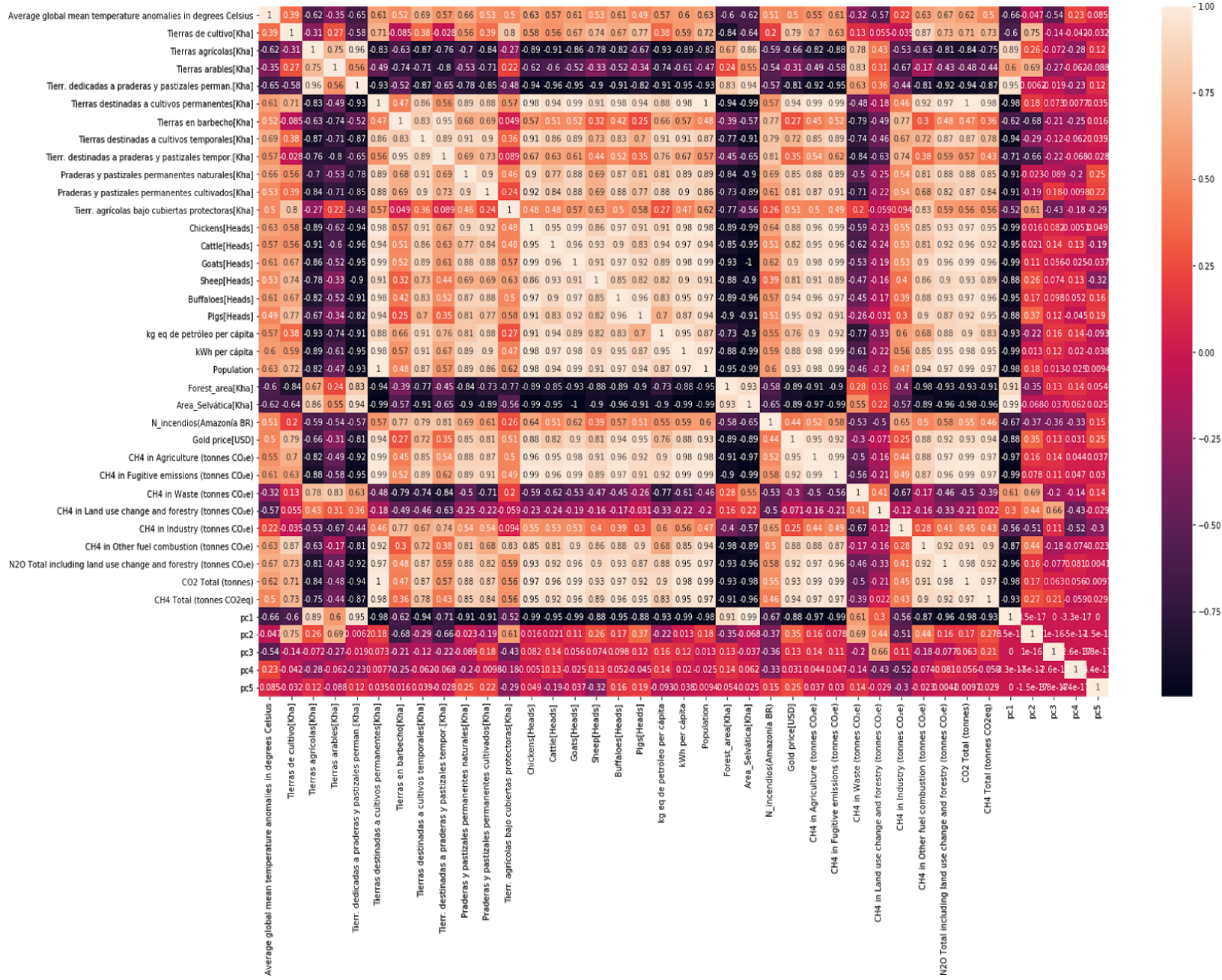
Luego de analizar las correlaciones y graficar, se procede a aplicar un **análisis de componentes principales**(PCA) con el objetivo de reducir la dimensionalidad del Dataset y obtener características que tengan correlación nulas, es decir, que son perpendiculares. Para esto se deduce a partir de la investigación que deberían haber al menos un componente principal por cada causa principal, teniendo a la agricultura, ganadería, energía eléctrica, combustibles fósiles y deforestación. Como los entes anteriores suman un total de 5, se decidió aplicar PCA para 5 componentes principales obteniendo:



Los gráficos nos indican la varianza de cada componente y como la suma de estas logran

representar el **0.959005285328359** de la varianza de los datos. Por lo tanto, bajo este criterio incluir más componentes principales no alterará los resultados de forma significativa, por lo que se decide mantener el número de componentes.

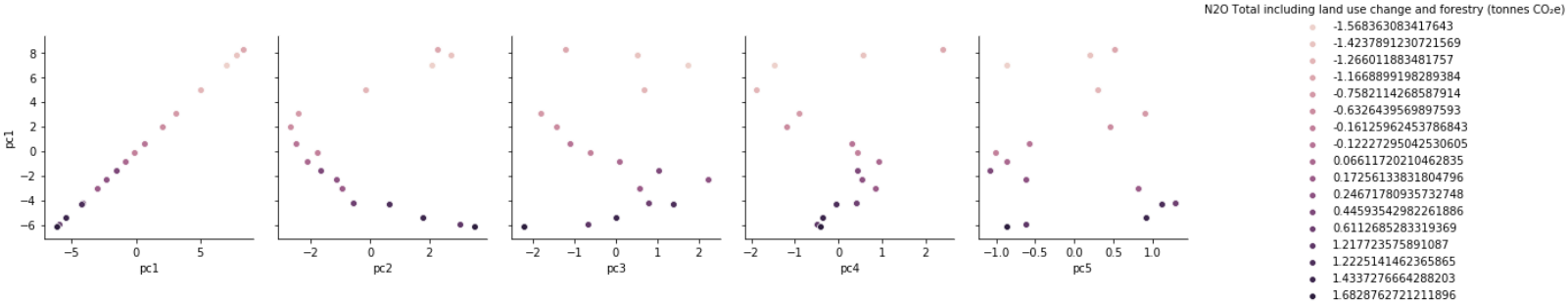
Por otro lado también obtenemos la matriz de correlación de los PC(Principal Components), para verificar el buen funcionamiento de PCA.

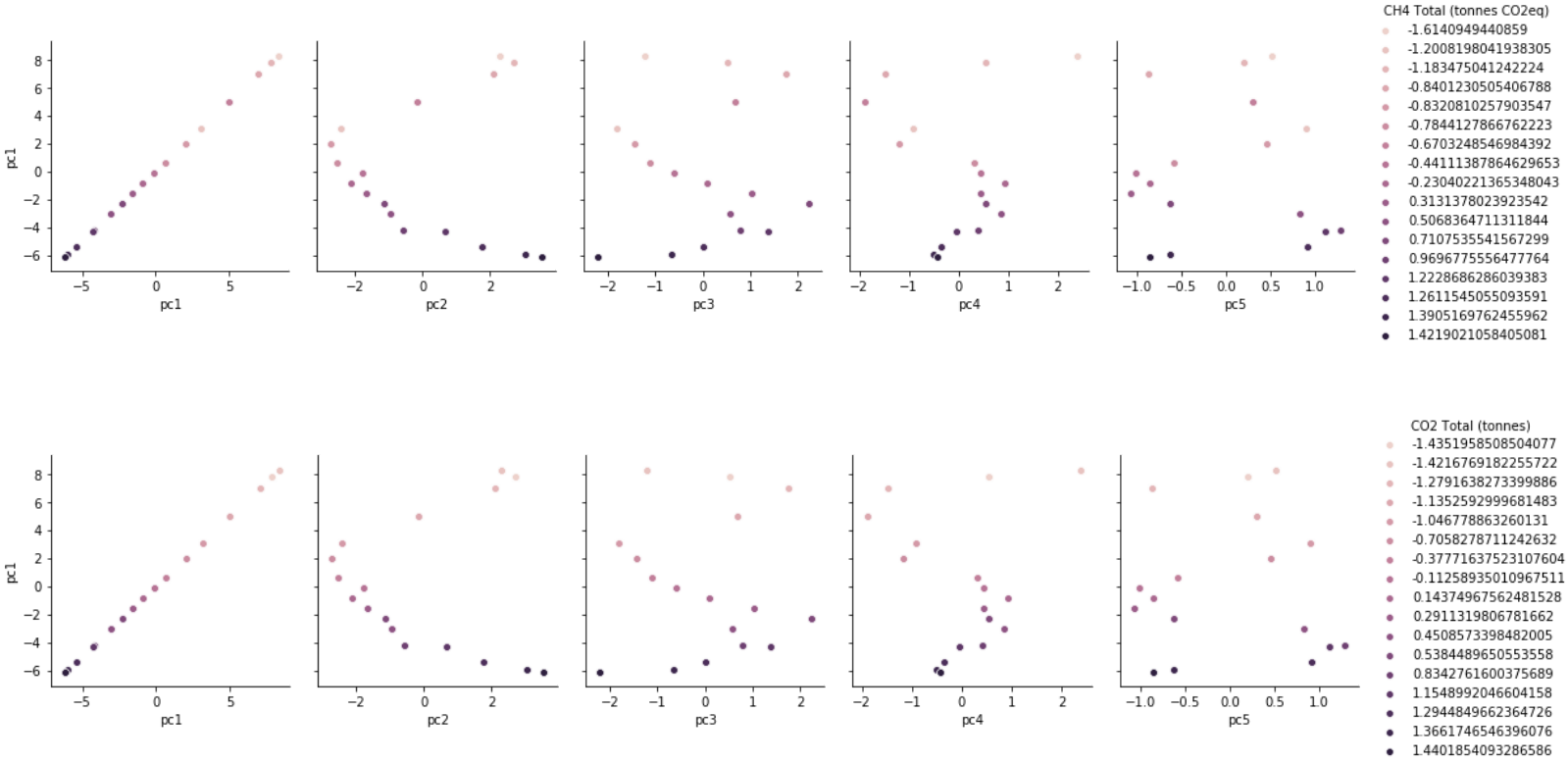


Como se observa, todos los componentes son ortogonales, es decir, tienen correlación 0 o muy cercano a 0. Por otro lado, se puede identificar una correlación negativa bastante alta por parte de la primera componente principal con las emisiones de **GEI**, que a su vez presenta **correlaciones negativas** con la **ganadería**, la **agricultura** y la **quema de combustibles fósiles**, también cabe mencionar que tiene alta **correlación** positiva con la **deforestación**, **tierras dedicadas a praderas y pastizales** y **tierras agrícolas**, esta última puede parecer un mal entendido, pero la definición de este tipo de tierra aglomera las tierras para pastoreo, cultivables(anuales y barbecho), plantación de árboles(frutales o para madera), etc. En fin, las **tierras agrícolas** tienen una alta variabilidad por la misma causa, por lo tanto no se puede saber con claridad los factores que la varían y la correlacionan con **pc1**, pero se presume al ver la matriz de correlaciones que se debe a la disminución de bosques, praderas y pastizales. Por otro lado, **pc2** está correlacionada con las **tierras arables y de cultivos** junto a la emisión de **gas metano** por parte de la **basura orgánica**; **pc3** por su parte está más **correlacionada positivamente** con la **absorción de gas metano por la silvicultura y el cambio de tierras** y **negativamente** con las **variaciones de temperatura**(menos que **pc1** pero más que las demás componentes); **pc4** está **correlacionada negativamente** con el **gas metano** emitido por las **industrias** y la **absorbida** por el **cambio de tierras y silvicultura**; **pc5** tiene bajas correlaciones, pero entre las más altas con apenas 0.25 de correlación estarían las **praderas y pastizales permanentes naturales y cultivados.**

Luego se procede a graficar las **componentes versus componentes** para identificar posibles grupos de interés que puedan permitirnos **clusterizar** y reconocer la existencia de entidades que participan con mayor o menor impacto en la emisión de gases de efecto invernadero. De estos grupos claramente podemos identificar unos de bajo, mediano y alto impacto en emisiones de **GEI**.

Por otra parte se decide incluir solo la fila que muestra la interacción de la primera componente con las demás, ya que aquí se encuentran los gráficos visualmente más aptos para analizar y más sencillos de clusterizar. Los gráficos se muestran con una variable de interés siendo esta cada uno de nuestros GEI.





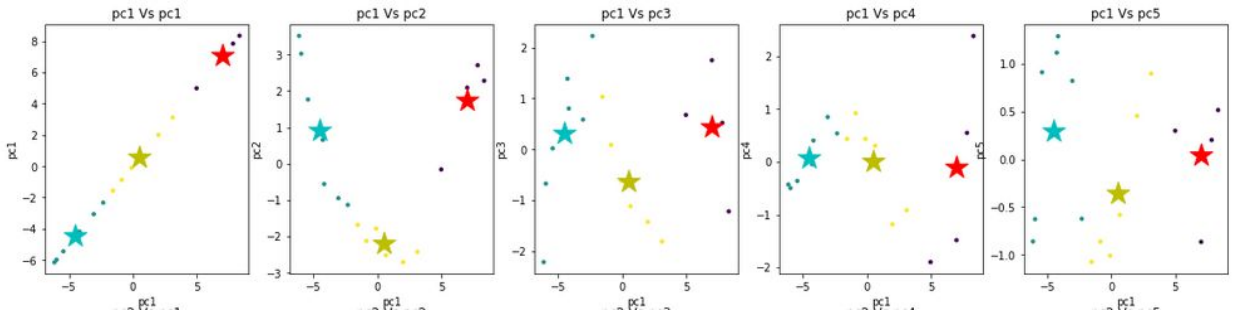
Para este punto, teniendo en cuenta la correlación de **pc1** con los **GEI** podemos observar que mientras más bajo es **pc1** más grande es la emisión de **GEI** (**pc5 vs pc1**). Luego se comienza la clusterización, utilizando en este caso dos algoritmos, **K-means** y **AgglomerativeClustering**.

Kmeans

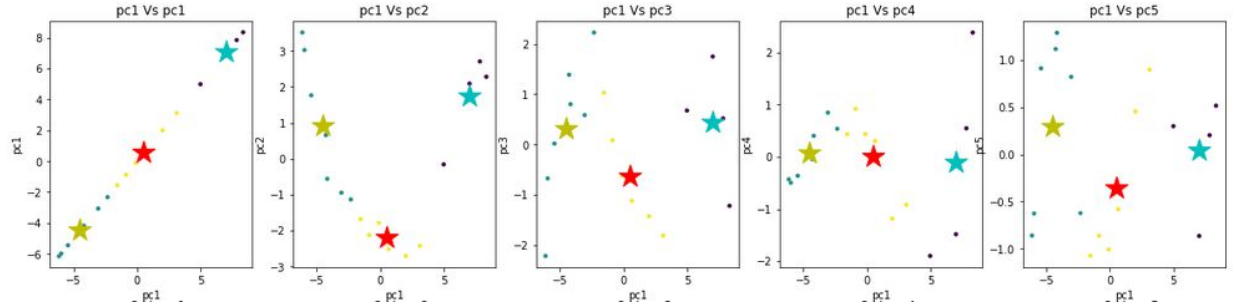
Para el caso de **K-means** se decide operar con dos variaciones del algoritmo, utilizando dos estilos del **algoritmo esperanza-maximización**(EM), **"full"** y **"elkan"** que trabajan internamente en el algoritmo o método **K-means** que provee la biblioteca **sklearn**. Además por inspección visual se eligen **tres grupos** y para asegurar el mejor centro de cada cluster se escogen **cien iteraciones** para cada ejecución.

Luego, se analizan los gráficos **pcx vs pc1** ya que contienen los mejores conjuntos de parámetros.

Utilizando el estilo **"full"** de expectation–maximization (EM) algorithm.



Utilizando el estilo **"elkan"** de expectation–maximization (EM) algorithm.



Debido a la cantidad de datos ambos estilos convergen al mismo resultado incluso después de **cien iteraciones**.

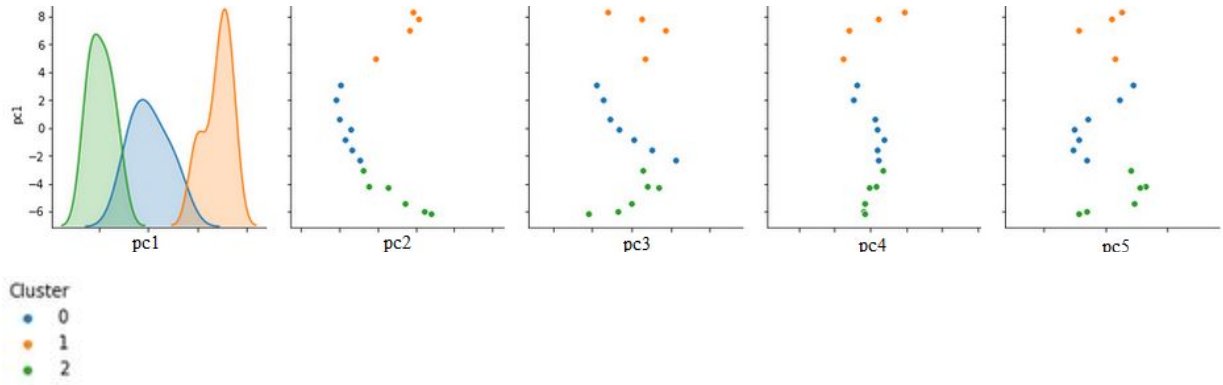
Ahora se imprimen los datos y el cluster al que pertenecen.

```
        pc1       pc5
0  8.339393  0.515365
1  7.835072  0.202012
2  7.015974 -0.865770
3  4.982129  0.298929
 Size:  4
         pc1       pc5
10 -2.326829 -0.623767
11 -3.063014  0.821591
12 -4.175396  1.288323
13 -4.282396  1.116430
14 -5.436532  0.911841
15 -5.974122 -0.627216
16 -6.141834 -0.860786
 Size:  7
         pc1       pc5
4  3.114370  0.896632
5  1.999369  0.454030
6  0.647676 -0.582214
7 -0.107914 -1.009124
8 -0.871915 -0.861990
9 -1.554032 -1.074286
 Size:  6
```

Como se muestra en la imagen, se obtienen tres grupos como se había decidido antes de la ejecución de **K-means**, los cuales además presentan distintos tamaños. El más pequeño de cuatro elementos(Derecha) y los dos más grandes de 6 (Centro)y 7 elementos(Izquierda). Estos grupos corresponden a las gráfica **pc5 vs pc1**, la cual presentaba la mejor visualización de los clusters.

Agglomerative Clustering

Por otro lado, con **Agglomerative Clustering** eligiendo nuevamente **tres grupos**, se obtienen los siguientes resultados.



Nuevamente se decide por mostrar solo la sección que representa las gráficas **pc1 vs pcx**(ejes invertidos con las gráficas anteriores, debido al uso de otra herramienta para graficar), ya que estas muestran buenos resultados de clustering, pero claramente **pc1 vs pc5** es superior a los demás.

Finalmente tras todo el trabajo de clustering se puede concluir que **Agglomerative Clustering** hizo un mejor trabajo con respecto a K-means, ya que este algoritmo logró identificar de mejor manera los grupos dejando **4 arriba, 7 al centro y 6 abajo**, es decir que en el gráfico de **K-means** uno de los elementos que pertenecía al grupo de la izquierda en este algoritmo pasó a ser parte del cluster del centro, dejando con más claridad la diferencia de unos de otros. Tras esto se puede decir que existen tres grupos de interés que impactan de forma particular en la emisión de gases de efecto invernadero. Siendo el **grupo** más **pequeño**(Arriba) el de **menor impacto**, el más **grande**(Centro) de **mediano impacto** y el de **seis elementos**(abajo) el de **mayor impacto** en la **emisión de GEI**.

Esto último nos permite identificar a todas las entidades altamente correlacionadas negativamente con **pc1**, como entes que generan un alto impacto en la emisión de **GEI**.

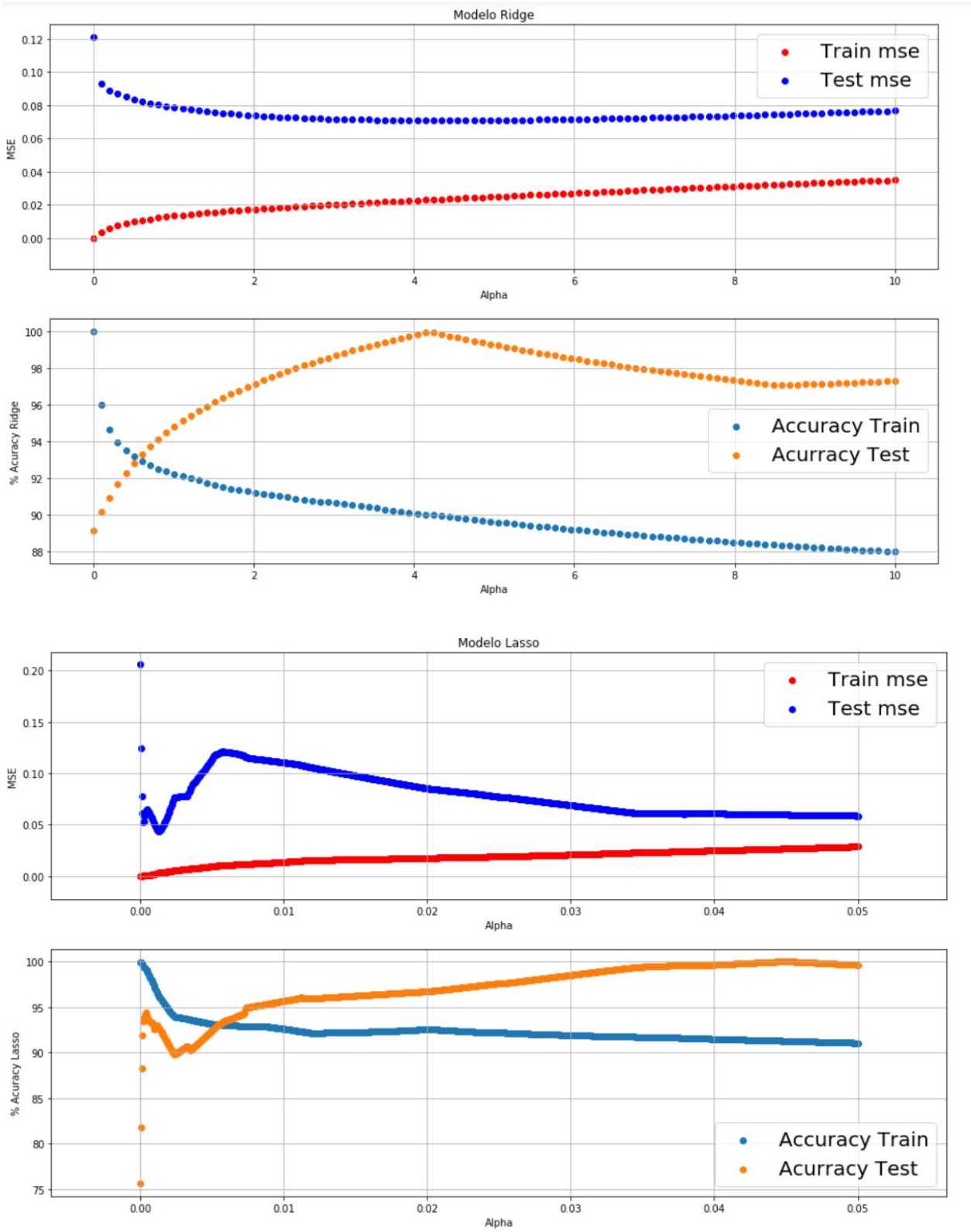
Regresión

Para el análisis de regresión se utilizaron los modelos de **Ridge**, **Lasso** y **Random Forest + Random Search**, con el objetivo de(a partir de un buen modelo) obtener los coeficientes de cada característica para ver su contribución en la predicción del valor de los **GEI**. En el caso de Random Forest se calcula la contribución media de cada característica en los árboles de decisión.

Para esto se seleccionó del dataset las características con **mayor correlación absoluta** (mayor a 0.5) con las variables **target** (CH4, N2O y CO2), luego se realizó el split del dataset con el fin de generar un test set de tamaño de 20% del dataset original, y así poder validar nuestros modelos.

Los resultados obtenidos son los siguientes:

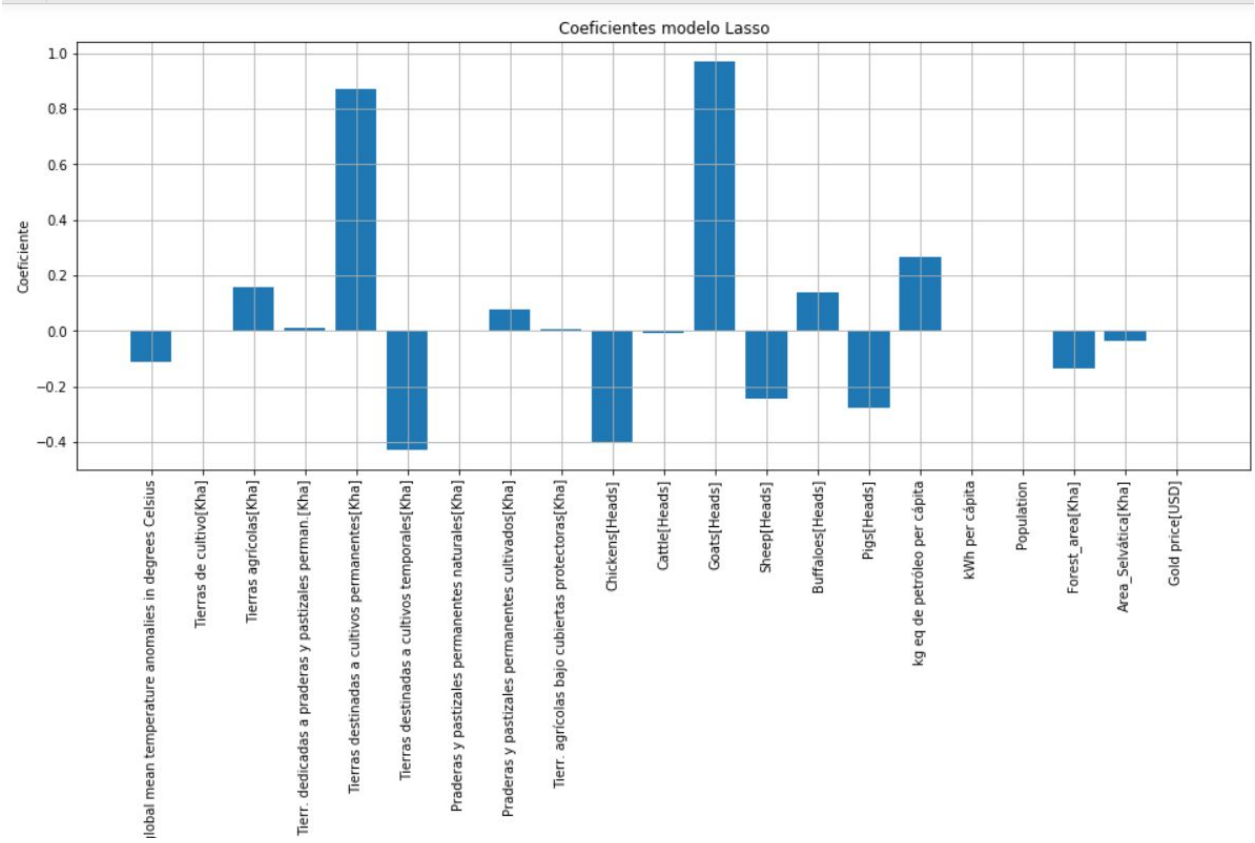
Metano (CH4):

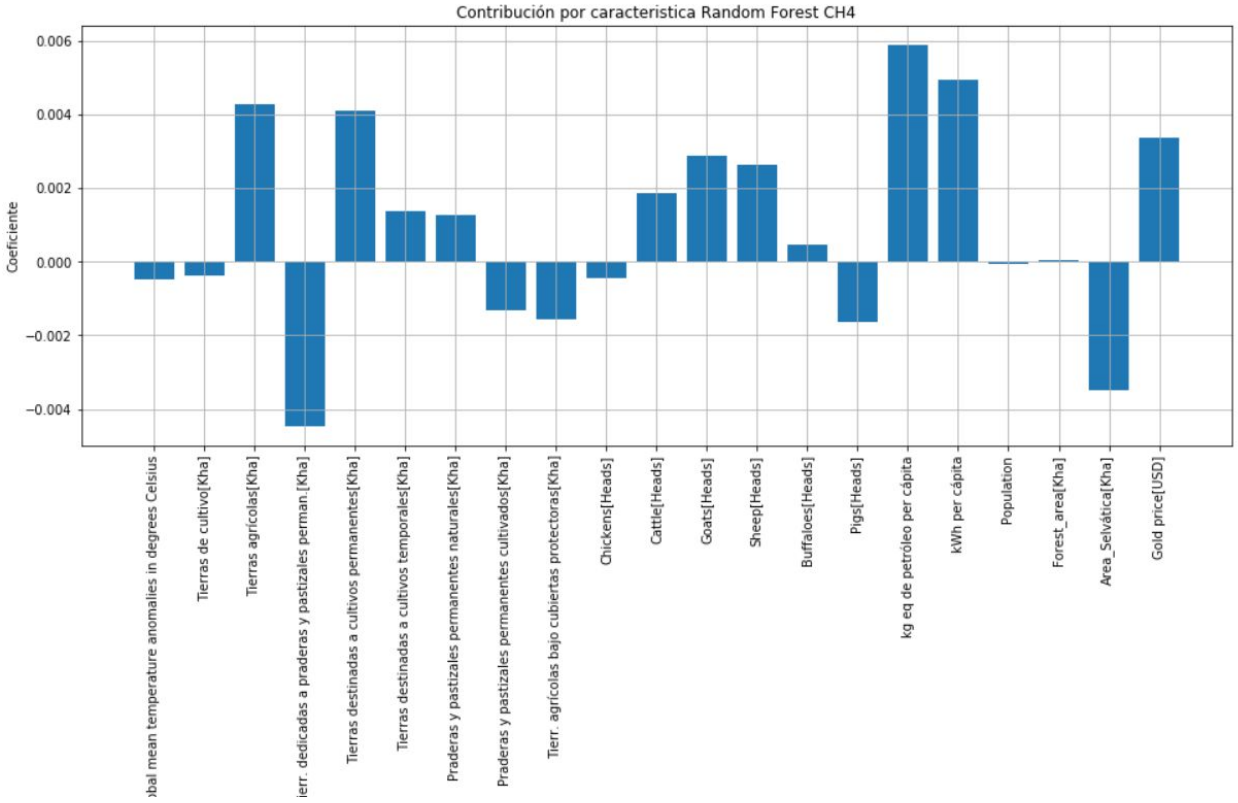


Random Forest

```
Model Performance
MSE train:  0.018094947579764712
MSE test:  0.06598480256627903
Accuracy_train = 96.66%.
Accuracy_test = 98.27%.
```

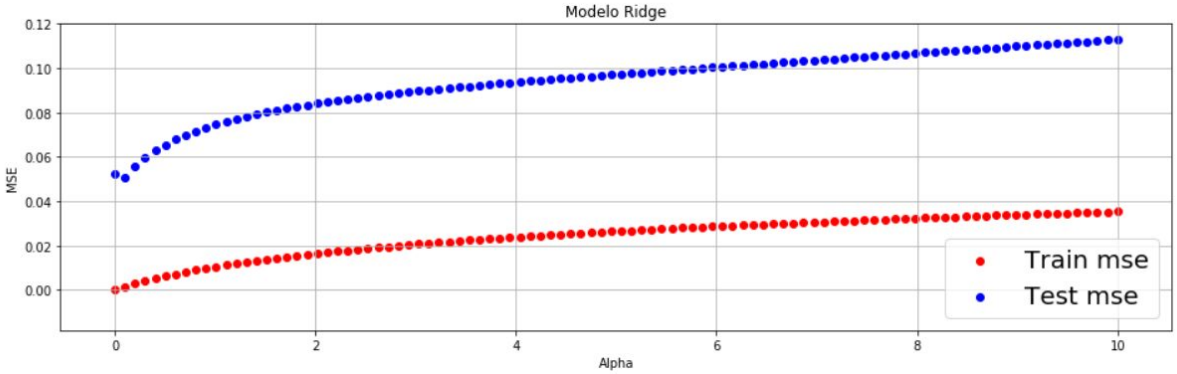
Para el caso de **CH4**, **Random Forest** y **Lasso** (con regularización 0.001 aprox.) se desempeñan de manera muy similar por lo que compararemos ambos coeficientes y contribuciones al modelo.

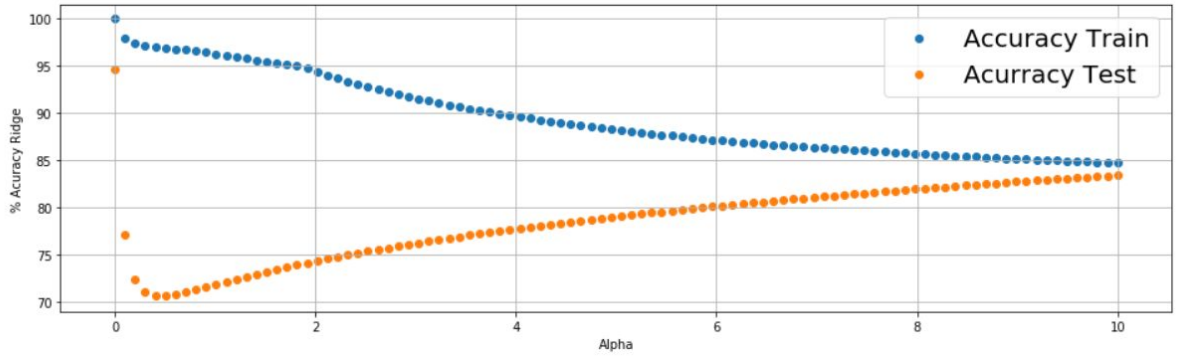


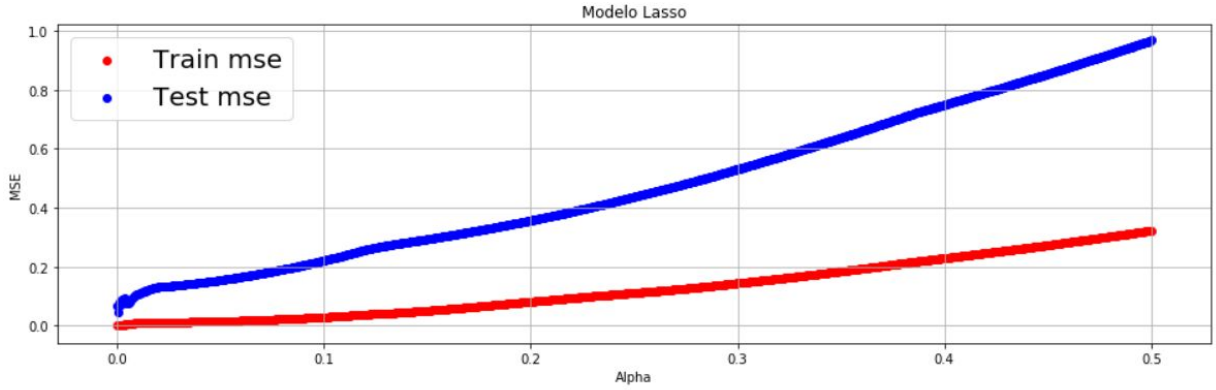


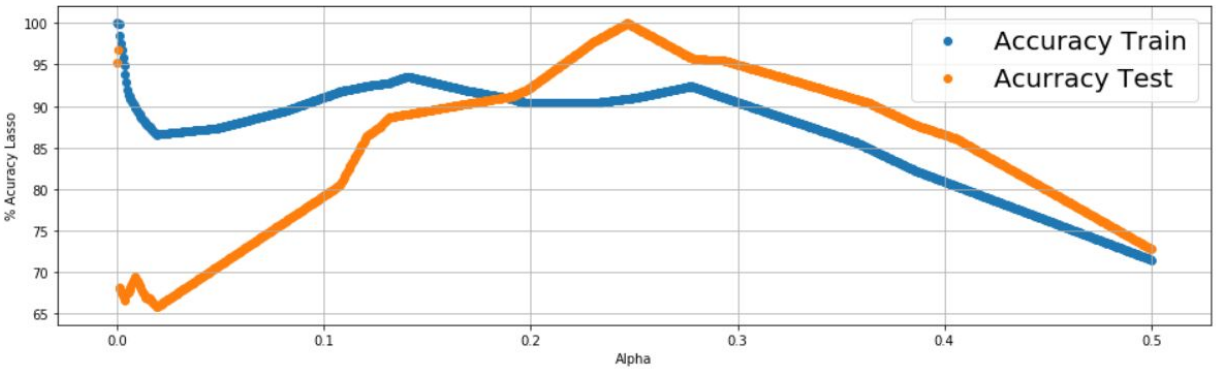
Para **Lasso**, los grandes contribuidores con la contaminación de metano son tierras destinadas a cultivos permanentes y la industria ganadera, específicamente de las cabras (goats), el resto de medidas son pequeñas y se podría decir que su valor negativo es para poder ajustar el modelo. Para la contribución de características de random forest, los mayores impactos se pueden ver en "**tierras dedicadas a praderas y pastizales perm.**", con una contribución negativa y para "**Kg. eq de petróleo per cápita**" y "**kWh per cápita**", con una contribución positiva. Ambas hacen sentido con respecto al significado de cada característica. En el **área ganadera**, la contribución de cada característica pero no menor, y en totalidad puede ser un aporte significante.

Oxido Nitroso (N2o):





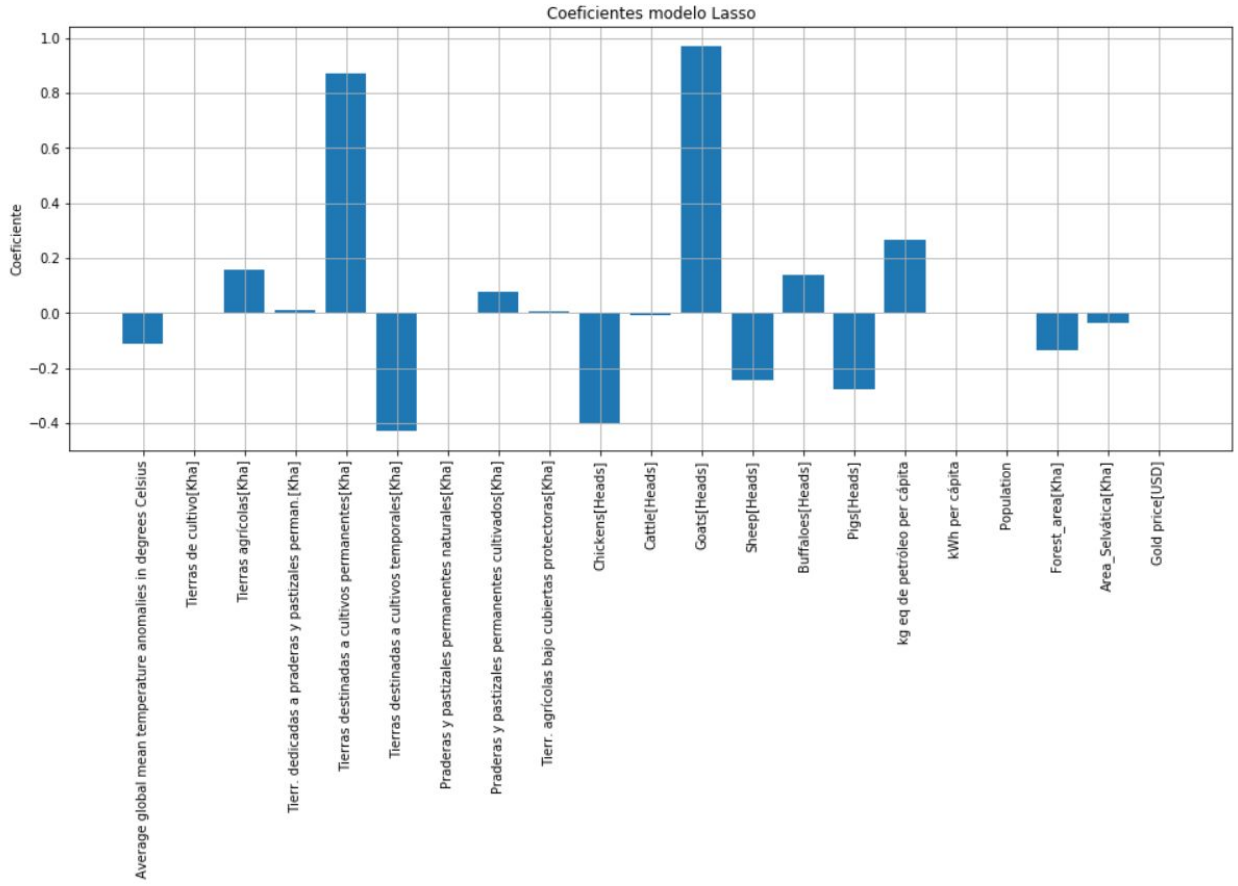




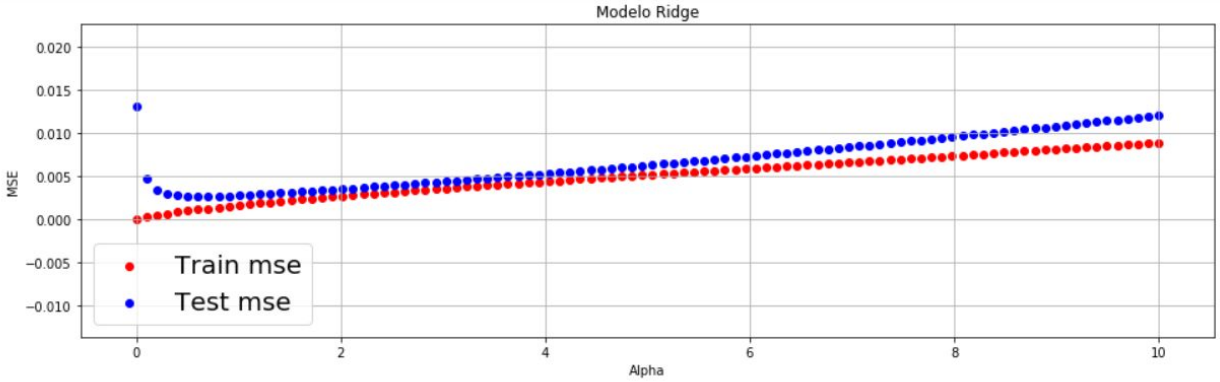
Random Forest

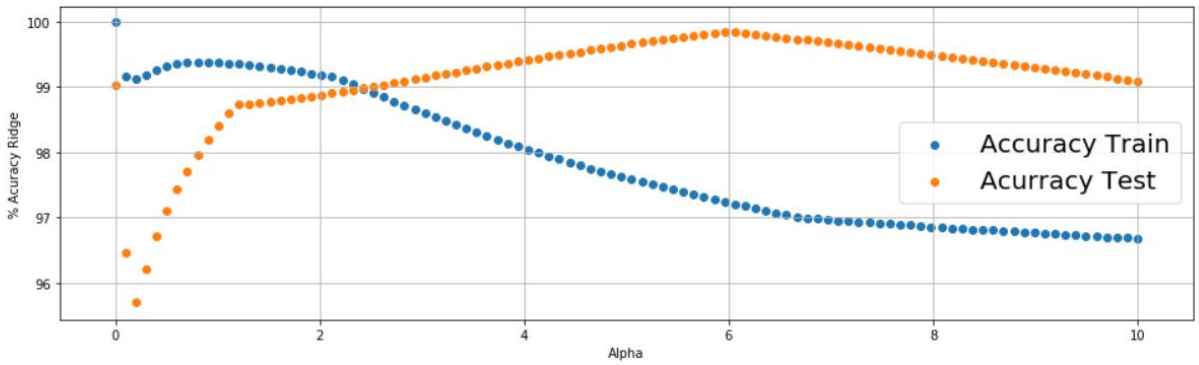
```
Model Performance
MSE train:  0.015917948710635614
MSE test:  0.18019625504706246
Accuracy_train = 89.77%.
Accuracy_test = 97.69%.
```

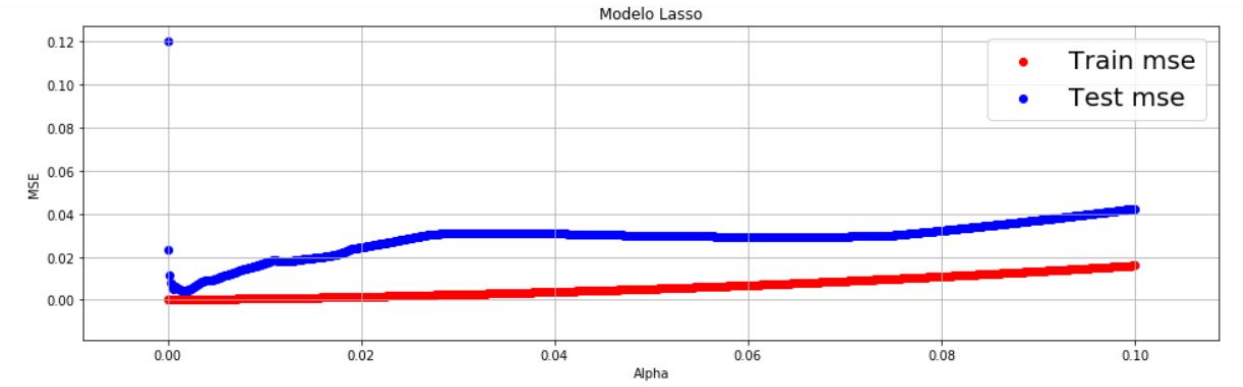
Para **óxido nitroso** (N2O) el mejor modelo obtenido es por **Lasso** (con regularización de 0.0007), por lo que se realizará el análisis de coeficientes de este modelo.

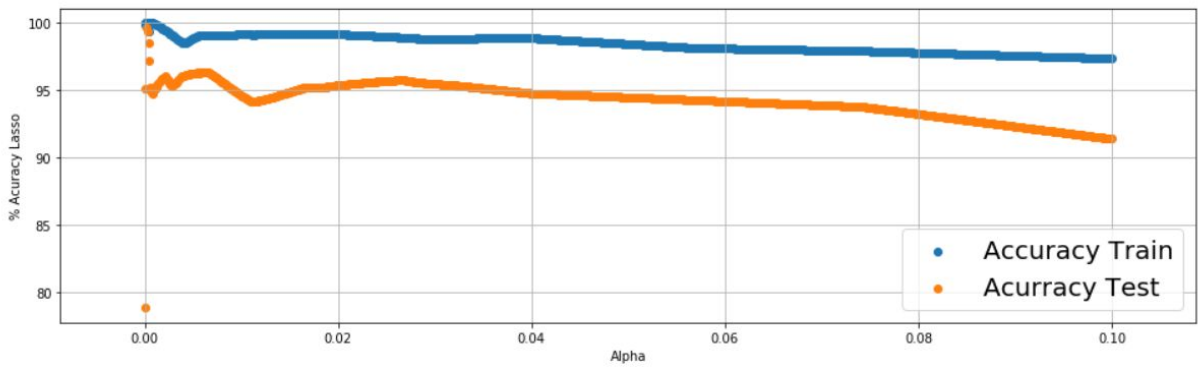


Dióxido de Carbono (CO2):





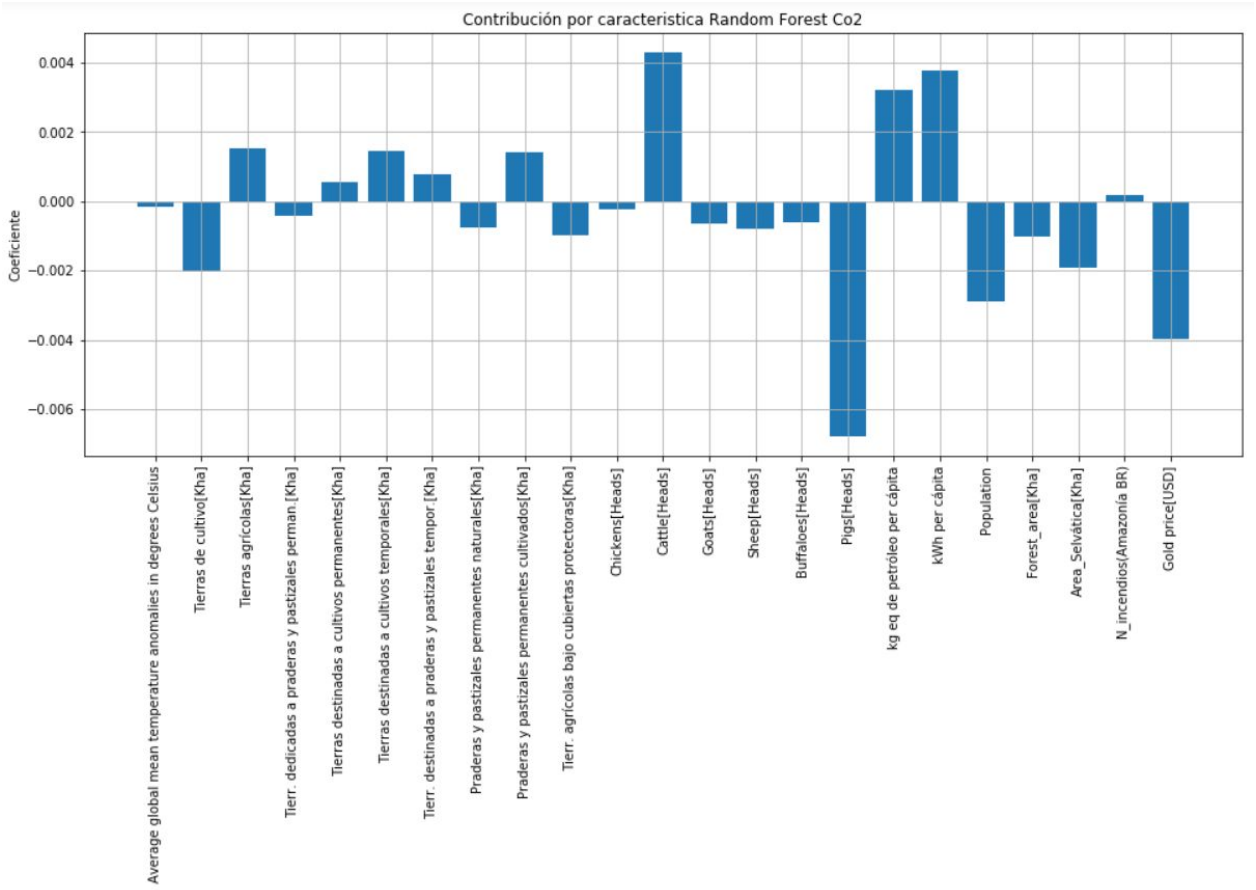




Random Forest

```
Model Performance
MSE train:  0.006764300814016238
MSE test:  0.04038285263992367
Accuracy_train = 97.92%.
Accuracy_test = 94.99%.
```

Para este caso se analiza las contribuciones de **Random Forest**:



Según lo anterior la **contribución negativa** más alta la tienen los **cerdos**(pigs), además como **contribuciones positivas** significativas están las **vacas** (cattle), **kg. eq de petróleo per cápita** y **kWh per cápita**.

Como se pudo verificar en las **regresiones**(para los distintos **GEI),** los mejores modelos fueron precedidos por las regresión de **Lasso** y **Random Forest**. Para el caso de **Lasso**, los coeficientes calculados por el modelo no nos explican de una manera coherente los resultados, aunque el modelo logra explicar con alta precisión en train y test la variable target. En el caso de **Random Forest** para **gas metano** (CH4), los resultados nos señalan que las **tierras de cultivos permanentes** tienen una **contribución negativa**, mientras que los **combustibles fósiles** y la

energía eléctrica contribuyen positivamente. Los resultados para **óxido nitroso**(N2O), obtenidos por **Lasso**, son idénticos a los de **CH4** , obtenidos por **Lasso** y se podría deber a que ambas variables target comparten las mismas características y con similar correlación. Por otro lado, para el caso de **Dióxido de Carbono**(CO2) nos indica que la industria ganadera de los cerdos (pigs) contribuye de una manera **negativa y significativa**, lo que contradice a nuestra **matriz de correlaciones** y a nuestros resultados obtenidos en **clustering**, pero también aunque en menor medida existe una contribución positiva por parte de las **vacas**(cattle), **combustibles fósiles** y **energía eléctrica**.

CONCLUSIÓN

De todo lo trabajado hasta ahora **no** podemos concluir que la **industria ganadera** es de las principales responsables de la emisión de **GEI**, esto se debe a falta de información en nuestro Dataset que nos impide hacer una acusación objetiva de esa magnitud. Por otro lado, si podemos concluir que la industria ganadera está más correlacionada con el **gas metano** que con el **óxido nitroso** y que está altamente correlacionada con los **cultivos permanentes** que a su vez tienen **correlación 1** con la **población mundial** y la **emisión de dióxido de carbono**. También podemos afirmar gracias a la **clusterización** y su análisis que la **industria ganadera** es un participante **activo** de la emisión de **GEI** y no se queda atrás en comparación a otras industrias. Sin embargo, para el análisis a través de los modelos de regresión no nos aportó mucha información ya que contradecía y afirmaba muchas veces lo obtenido a través de la matriz de correlaciones y lo investigado previamente, estas confusiones por parte de los modelos se debieron a la escasez de datos y la alta variabilidad de estos.

COMPLICACIONES Y DESAFÍOS

A lo largo de la investigación todo parecía posible, aunque hubieran varias entidades relacionadas al **cambio climático** no esperábamos tener tantas complicaciones con los Datasets. Estos además de ser difíciles de conseguir(incluso googleando en inglés), carecían de los datos o fechas que nos importaban, hablamos de datos que empezaban desde el 2010 hacia delante o que empezaban desde 1800 y terminaban el 2012, simplemente eran Datasets que no tenían lo que buscábamos incluso luego de acotar desde 1998-2014 por la misma situación.

También hubo al inicio una **subestimación del problema**, cuando identificamos las entidades más importantes de las emisiones nunca pensamos que debíamos, además, enfocarnos en las sub-entidades que dependían de estas más grandes, entonces cuando comenzamos a investigar más en cada industria o situación, nos encontrábamos con estos pequeños problemas que también

participaban en el proceso.

Por otro lado se nos presentó el problema de **no** tener un **volumen grande de datos**, el cual no nos permitió entrenar modelos de regresión que nos ayudaran a predecir las emisiones de gases, cada vez que hacíamos un modelo aunque muchos eran bastante parecidos, algunos no tenían ningún sentido y esto se debía precisamente a que las características que seleccionaba el modelo o eran muy distintas o muy parecidas, lo que alteraba bruscamente los resultados. ¿Solución?, Más datos y más características, es decir, entes que interactúan en el sistema.

Las dificultades que entregan los **datos heterogéneos** son su alta varianza o dispersión. Esto junto a una pequeña cantidad de datos puede llevar a resultados no deseados ya que no existe un patrón o convergencia clara.

Durante este proceso se decidió en la mayoría de los casos eliminar a los países a quienes les faltara el 30% de los datos, luego de esto quedaban una cantidad mínima de países que aún tenían NaNs, a los cuales se les **imputó** el promedio de su población. Este proceso se hizo cuando analizábamos cada uno de los Datasets que convergieron al Dataset final, en los cuales la cantidad de datos era bastante alta como para no percibir esas imputaciones, en el caso de pocos datos esas prácticas claramente pueden terminar por contaminar los datos.

Finalmente, de esta experiencia, pudimos darnos cuenta del **pobre proceso de análisis del problema** que hicimos al inicio, este fue nuestro más grande error. A medida que avanzabamos el problema se agrandaba más y más y ya estábamos bastante sumergidos en la temática como para retroceder, todo nuestro tiempo invertido en investigación se multiplicó por dos(24 hrs mínimo) y el tiempo en búsqueda de Datasets se hizo un infierno(30 hrs mínimo), para luego recién poder iniciar el proceso de análisis de datos, el cual fue más expedito pero no menos problemático. Lo único que nos dió **satisfacción** fue encontrar Datasets que tuvieran estructuras parecidas en los cuales podíamos utilizar algunos de los métodos antes utilizados para **pre-procesar** los datos e incluirlos en el Dataset final. Por lo antes mencionado estamos seguros que si tuviéramos que hacer un trabajo de este calibre y estilo a futuro, deberíamos adoptar métodos más sofisticados, como los aprendidos en ingeniería de software para identificar requisitos, ordenar tareas por prioridad e identificar el **objetivo** principal desde el comienzo.

REFERENCIAS(linkografia)

1. EL TRABAJO DE LA FAO SOBRE EL CAMBIO CLIMÁTICO
2. Resultados | Modelo de Evaluación Ambiental de la Ganadería Mundial (GLEAM) | Organización de las Naciones Unidas para la Alimentación y la Agricultura
3. La ganadería y el medio ambiente | FAO | Organización de las Naciones Unidas para la Alimentación y la Agricultura
4. La ganadería amenaza el medio ambiente
5. http://www.fao.org/faostat/es/#data/EM/visualize (Distribución de emisiones)
6. ¿Qué son las PM2,5 y cómo afectan a nuestra salud?
7. Indicators (BancoMundial)
8. Mining drives extensive deforestation in the Brazilian Amazon
9. Minería ilegal: la peor devastación en la historia de la Amazonía
10. La fiebre del oro en la Amazonía destruye la selva tropical
11. La pérdida de cobertura arbórea mundial ascendió al 51% en 2016
12. Global fire data
13. Los 30 países con más área forestal (2020) • Libretilla
14. ¿Cómo afecta el consumo de energía al medio ambiente?
15. Redalyc.La generación de energía eléctrica y el ambiente
16. Energía renovable para abastecer a todo el planeta
17. EMISIONES DE GASES
18. World Population Growth
19. ¿Amazonía o Amazonas?: cuál es la forma correcta para hablar del incendio que afecta a esta región
20. Global Carbon Budget 2019
21. Cultivo bajo cubierta
22. Big Data y la lucha contra el cambio climático | OpenMind
23. FUGITIVE EMISSIONS
24. CH4 EMISSIONS FROM SOLID WASTE DISPOSAL
25. Sector uso de la tierra, cambio de uso de la tierra y silvicultura
26. Gases de efecto invernadero
27. Gases de efecto invernadero y el cambio climático (Instituto de Hidrología, Meteorología y Estudios Ambientales - IDEAM)
28. Interpreting random forests
29. CO_2 and Greenhouse Gas Emissions
30. Estadísticos de dispersión