Experiencia en el Data Challenge

Análisis de datos de la industria cafetalera

Angel Luis Robles Fernández, Carlos Manuel Rodríguez Martínez

24 de noviembre de 2017

Facultad de Física UV

Tabla de contenidos

Introducción: Industria cafetalera y sus problemas

Intro

- El café es un cultivo prioritario para el país
- Emplea 500 mil productores
- Involucra 897 millones de dólares / año
- Vincula 3 millones de personas en México

Intro

- Veracruz produce el 26 %
- El café tiene un valor agregado
- Taza de excelencia

Intro

¿Qué características se buscan?

Aroma, cuerpo, acidez y fineza

- Se huelen, catan y puntúan los mejores cafés
- Se crea un índice de café (Índice de taza de excelencia ITE) de 0 a 100

Intro

¿Qué características se buscan? Aroma, cuerpo, acidez y fineza

- Los mejores cafés se subastan
- Existe una enorme demanda de cafés singulares
- Alcanzan precios de US$ 100.49/lb
- Ganador 2017: located at Naolinco, Veracruz, Mexico at an altitude of 1.240m.
- http://www.allianceforcoffeeexcellence.org/en/cup-of-excellence/country-programs/mexico-program/2017/auction-results/

Intro

El problema: Modelar la Relación del ITE y las condiciones ambientales

Datos disponibles

Base de datos

Las bases de datos de empresas, organizaciones o ciencias sociales suelen ser extremadamente desordenadas. Esta no es la excepción. Tenemos una base de datos de 1158 muestras de cataciones de café. Se miden parámetros de calidad.

ESTADO	REGIÓN	LOCALIDAD	MUNICIPIO2	OL_1	OL_2	OL_3	OL_4	OL_5	OL_6	OL_7	OL_9	OL_10	OL_11	OL_12	OL_D	OL_TOT
VERACRUZ	0.	EL OLVIDO	ZENTLA													
0.	0.	0.	0.													
VERACRUZ	COATEPEC	LA PERLA	IXHUACAN DE LOS REYES	2085.	8.08333	8.08333	7.83333	8.	7.66667	10.	7.91667	10.	10.	8.16667	0.	85.75
VERACRUZ	COATEPEC	LA PERLA	IXHUACAN DE LOS REYES	2084.	7.5	7.5	7.33333	7.58333	7.58333	10.	7.58333	10.	10.	7.66667	0.	82.75
VERACRUZ	COATEPEC	LA PERLA	IXHUACAN DE LOS REYES	2083.	7.91667	8.33333	7.83333	7.58333	7.41667	10.	7.91667	10.	10.	8.08333	0.	85.0833
CDMEX		OZOMATLÁN	HUAUCHINANGO	2082.	8.	8.25	8.	7.91667	7.58333	10.	7.91667	10.	10.	8.16667	0.	85.8333
CDMEX		EJIDO NACTANCA	XILIPETEC PUEBLA	2081.	7.5	7.58333	7.33333	7.66667	7.58333	10.	7.41667	10.	10.	7.41667	0.	82.5
CDMEX		OZOMATLÁN	HUAUCHINANGO PUEBLA	2080.	7.33333	7.16667	7.08333	7.5	7.25	10.	7.08333	10.	10.	7.	0.	80.4167
CDMEX		LAS PILAS	XICOTEPEC, PUEBLA	2079.	6.83333	6.91667	6.91667	7.5	7.25	10.	6.91667	10.	10.	6.91667	0.	79.25
CDMEX		EL NZACATAL	TLACUILOTEPEC, PUEBLA	2078.	7.41667	7.5	7.25	7.33333	7.41667	10.	7.33333	10.	10.	7.25	0.	81.5
CDMEX		TLAPEHUALITA	TLACUILOTEPEC, PUEBLA	2077.	7.	7.08333	6.83333	7.16667	7.16667	10.	6.91667	10.	10.	6.83333	0.	79.

Figura 1: Base de datos de cataciones.

Datos disponibles: (ESTADO, REGIÓN, LOCALIDAD, MUNICIPIO2, OL_1, OL_2, OL_3, OL_4, OL_5, OL_6, OL_7, OL_9, OL_10, OL_11, OL_12, OL_D, OL_TOT)

Base de datos

Los datos bioclimáticos representan factores de temperatura, humedad, precipitación, etc de un área geográfica. La base de datos contiene 5156 muestras.

POINT_X	POINT_Y	bio_1	bio_2	bio_3	bio_4	bio_5	bio_6	bio_7	bio_8	bio_9	bio_10	bio_11	bio_12
−96.9415	19.9696	1.58824	0.354798	−0.422634	1.62109	1.51755	1.50011	0.995507	0.90564	1.80846	1.69151	1.2868	−0.182831
−96.9259	18.7633	−0.724042	0.533431	1.40813	−1.26615	−0.780754	−0.710341	−0.616741	−0.893928	−0.632262	−0.852535	−0.463882	1.51992
−97.0465	19.9963	1.16232	0.622747	−0.535068	1.97337	1.28772	0.868555	1.53292	1.23985	1.40302	1.37882	0.719944	−0.283234
−96.9335	19.454	−0.940484	0.00825129	−0.0110839	−0.284995	−0.734788	−1.02612	0.0281578	−0.782355	−1.08027	−0.910332	−0.980613	−0.473795
−96.8184	19.573	−0.107934	−0.419276	−0.621136	0.233286	−0.0452975	−0.0787825	0.0281578	−0.0368195	−0.43816	−0.0784528	−0.175086	−1.59872
−96.9962	18.9531	−0.610464	0.444114	1.04992	−0.947262	−0.642856	−0.647185	−0.401775	−0.713971	−0.606923	−0.712736	−0.423904	0.32533
−96.9439	19.4422	−0.867086	−0.0620108	0.202874	−0.354874	−0.734788	−0.836653	−0.294292	−0.713971	−0.995633	−0.860933	−0.871419	−0.447157
−96.9396	19.0667	−0.380629	0.494132	0.900812	−0.744407	−0.36706	−0.394562	−0.186809	−0.465631	−0.446776	−0.457343	−0.215064	−0.0250553
−96.8221	18.8253	1.13982	1.37777	0.563207	−0.057359	1.33368	0.994867	1.42544	1.14575	0.820209	0.99153	1.26651	0.343772
−97.0692	20.0063	1.32037	0.812097	−0.28535	1.84333	1.42562	1.05802	1.53292	1.40283	1.55506	1.48552	0.928784	−0.680748

Figura 2: Base de datos de características bioclimáticas.

Datos disponibles: POINT_X, POINT_Y, bio_1, bio_2, bio_3, bio_4, bio_5, bio_6, bio_7, bio_8, bio_9, bio_10, bio_11, bio_12, bio_13, bio_14, bio_15, bio_16, bio_17, bio_18, bio_19. ¿¿Cómo se puede hacer uso de todos estos datos??

Modelos biológicos

Modelos

Existe una relación entre las condiciones ambientales favorables

Procesamiento de datos

Primero: Elegir un lenguaje de programación

Las discusiones entre lenguajes de programación **siempre** son una guerra religiosa. **Sólo opiniones personales**.

- **C/C++**
  - Lenguaje de muy bajo nivel.
  - Ideal para optimizar el rendimiento.
  - Largo tiempo de desarrollo, poco nivel de interacción.
- **Python**
  - Muy alto nivel.
  - Lenguaje interpretado, rendimiento intermedio.
  - Facilidad de interacción.
  - Poco consistente entre paquetes.

Primero: Elegir un lenguaje de programación

- **R**
  - Lenguaje libre, desarrollado por la comunidad y de alto nivel.
  - Interpretado y de rendimiento intermedio.
  - Interactivo.
  - Diseñado para ciencias.
- **Mathematica**
  - Extremadamente alto nivel.
  - Lenguaje interpretado y simbólico. Rendimiento bajo.
  - Diseño consistente.
  - Facilidad de integración con otros lenguajes.

Lo más aburrido: lidiar con los errores en la base de datos

El problema: En la vida real las bases de datos no son consistente. Cada operador introduce los datos de forma diferente. Método usual: Usar *regex*.

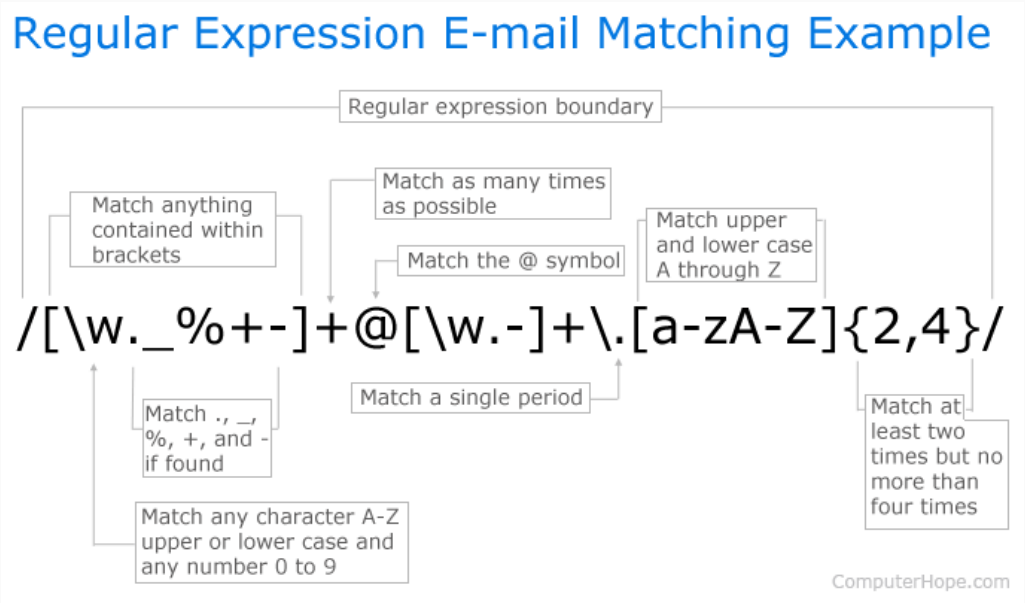

Figura 3: Sintaxis de regex.

Lo más aburrido: lidiar con los errores en la base de datos

El problema: En la vida real las bases de datos no son consistente. Cada operador introduce los datos de forma diferente. Mi método de preferencia: Usar patrones en el leguaje Wolfram.

ESTADO	REGIÓN	LOCALIDAD	MUNICIPIO2	OL_1	OL_2	OL_3
0.	0.	0.	0.	1955.	7.58333	7.58333
VERACRUZ	0.	CORDOBA	CORDOBA	1954.	7.83333	8.08333
VERACRUZ	0.	CORDOBA	CORDOBA	1953.	7.91667	8.16667
0.	0.	0.	0.	1952.	7.5	7.58333
0.	0.	0.	0.	1951.	7.66667	7.41667
0.	0.	0.	0.	1950.	7.5	7.58333
0.	0.	0.	0.	1949.	11.	11.125
0.	0.	0.	0.	1948.	7.5	7.83333
Veracruz	0.	0.	0.			
OAXACA	0.	0.	0.			
CHIAPAS	0.	0.	0.			
OAXACA	0.	0.	0.			
0.	0.	HUEHUETECPAN	0.			
0.	0.	0.	0.			
0.	0.	0.	0.			
0.		0.	0.	1924.	7.08333	6.91667
0.	0.	RODRIGUEZ CLARA	XICO			
0.	0.	0.	0.	1907.	7.66667	7.58333
0.	0.	0.	0.	1903.	7.58333	7.83333
0.	0.	0.	0.	1902.	7.66667	7.75

Figura 4: Base de datos de cataciones.

Lo más aburrido: lidiar con los errores en la base de datos

La solución: Corregir los errores y establecer "patrones" para cada situación.

Se formatean las entradas de forma consistente

```
rules = {
   (* Reglas generales *)
   <|a___, "ESTADO" → x_String, "REGIÓN" → _String, "LOCALIDAD" → y_String, "MUNICIPIO2" → _String, b___|> :> <|a, "POSFORMATEADA" → {x, y}, b|>,
   <|a___, "ESTADO" → x_String, "REGIÓN" → y_String, "LOCALIDAD" → 0, "MUNICIPIO2" → 0, b___|> :> <|a, "POSFORMATEADA" → {x, y}, b|>,
   <|a___, "ESTADO" → x_String, "REGIÓN" → 0, "LOCALIDAD" → y_String, "MUNICIPIO2" → _, b___|> :> <|a, "POSFORMATEADA" → {x, y}, b|>,
   <|a___, "ESTADO" → x_String, "REGIÓN" → 0, "LOCALIDAD" → 0, "MUNICIPIO2" → y_String, b___|> :> <|a, "POSFORMATEADA" → {x, y}, b|>,
   <|a___, "ESTADO" → x_String, "REGIÓN" → 0, "LOCALIDAD" → 0, "MUNICIPIO2" → 0, b___|> :> <|a, "POSFORMATEADA" → {x}, b|>,

   (* Reglas específicas *)
   <|a___, "ESTADO" → 0, "REGIÓN" → "Coatepec", "LOCALIDAD" → _, "MUNICIPIO2" → _, b___|> :> <|a, "POSFORMATEADA" → {"Veracruz", "Coatepec"}, b|>,
   <|a___, "ESTADO" → 0, "REGIÓN" → "Ixhuatlan", "LOCALIDAD" → _, "MUNICIPIO2" → _, b___|> :> <|a, "POSFORMATEADA" → {"Veracruz", "Ixhuatlan"}, b|>,
   <|a___, "ESTADO" → 0, "REGIÓN" → "Estado De Mexico", "LOCALIDAD" → _, "MUNICIPIO2" → _, b___|> :> <|a, "POSFORMATEADA" → {"Estado De Mexico"}, b|>,
   <|a___, "ESTADO" → 0, "REGIÓN" → "Huatusco", "LOCALIDAD" → _, "MUNICIPIO2" → _, b___|> :> <|a, "POSFORMATEADA" → {"Veracruz", "Huatusco"}, b|>,
   <|a___, "ESTADO" → 0, "REGIÓN" → "Zongolica", "LOCALIDAD" → _, "MUNICIPIO2" → _, b___|> :> <|a, "POSFORMATEADA" → {"Veracruz", "Zongolica"}, b|>,
   <|a___, "ESTADO" → 0, "REGIÓN" → "Cordoba", "LOCALIDAD" → _, "MUNICIPIO2" → _, b___|> :> <|a, "POSFORMATEADA" → {"Veracruz", "Cordoba"}, b|>,
   <|a___, "ESTADO" → 0, "REGIÓN" → "Colima", "LOCALIDAD" → _, "MUNICIPIO2" → _, b___|> :> <|a, "POSFORMATEADA" → {"Colima", "Colima"}, b|>,
   <|a___, "ESTADO" → 0, "REGIÓN" → "Atzalan", "LOCALIDAD" → _, "MUNICIPIO2" → _, b___|> :> <|a, "POSFORMATEADA" → {"Veracruz", "Atzalan"}, b|>,
   <|a___, "ESTADO" → 0, "REGIÓN" → "Misantla", "LOCALIDAD" → _, "MUNICIPIO2" → _, b___|> :> <|a, "POSFORMATEADA" → {"Veracruz", "Misantla"}, b|>
  };
```

Figura 5: Reglas de transformación para cada situación.

Resultados del preprocesamiento

A partir del preprocesamiento se obtienen dos bases de datos con las cuales podemos trabajar. Quedan 341 muestras útiles de datos de calidad de café.

Figura 6: Puntos geográficos de datos de cataciones.

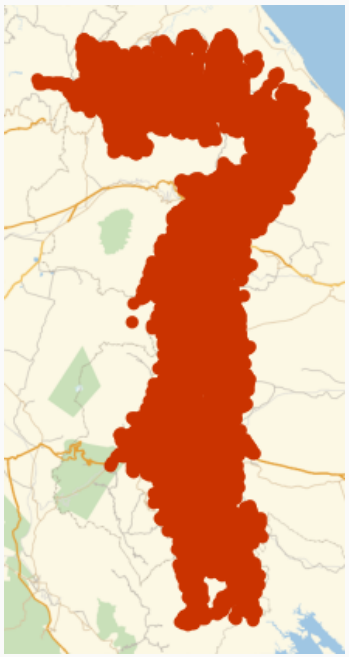

Figura 7: Puntos de datos de datos bioclimáticos.

Caracterización de datos bioclimáticos

Espacio ambiental y teoría del nicho ecológico

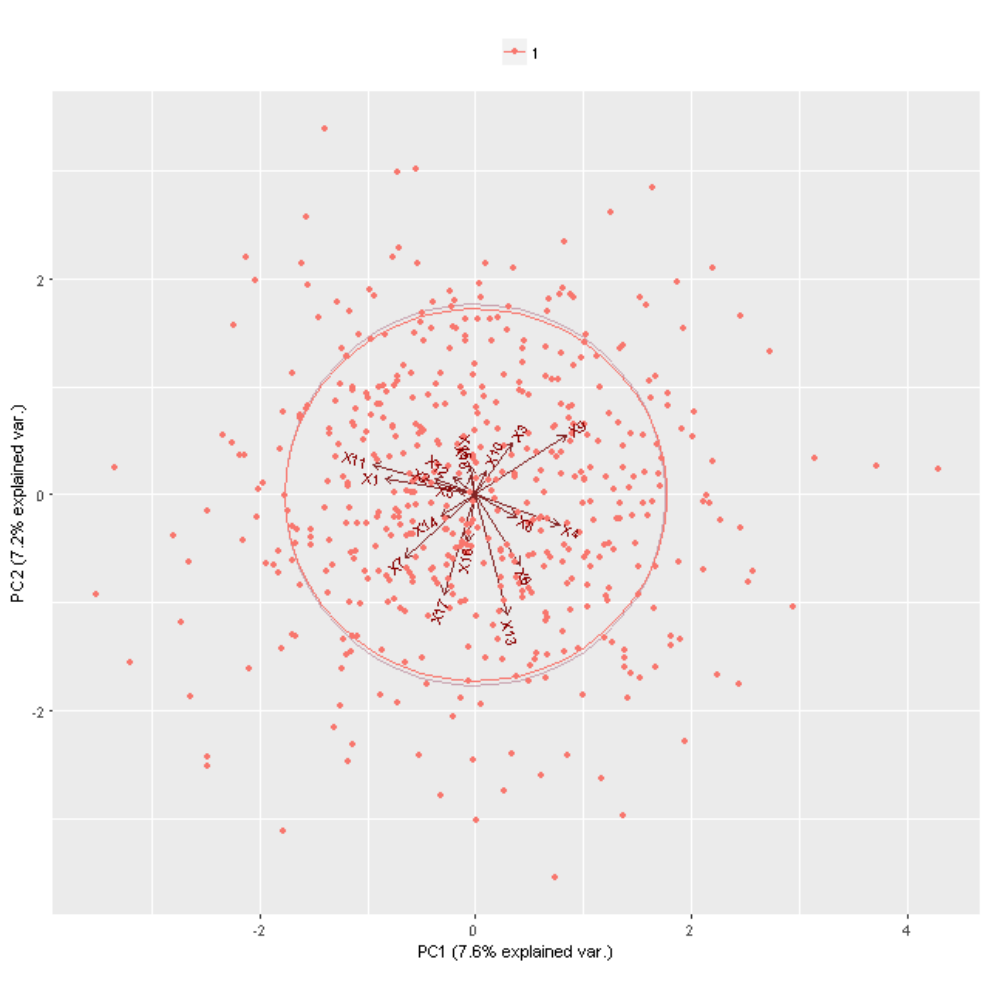

Figura 8: Espacio ambiental teórico (variables aleatorias normales).

Caracterización de datos bioclimáticos

Espacio ambiental y teoría del nicho ecológico

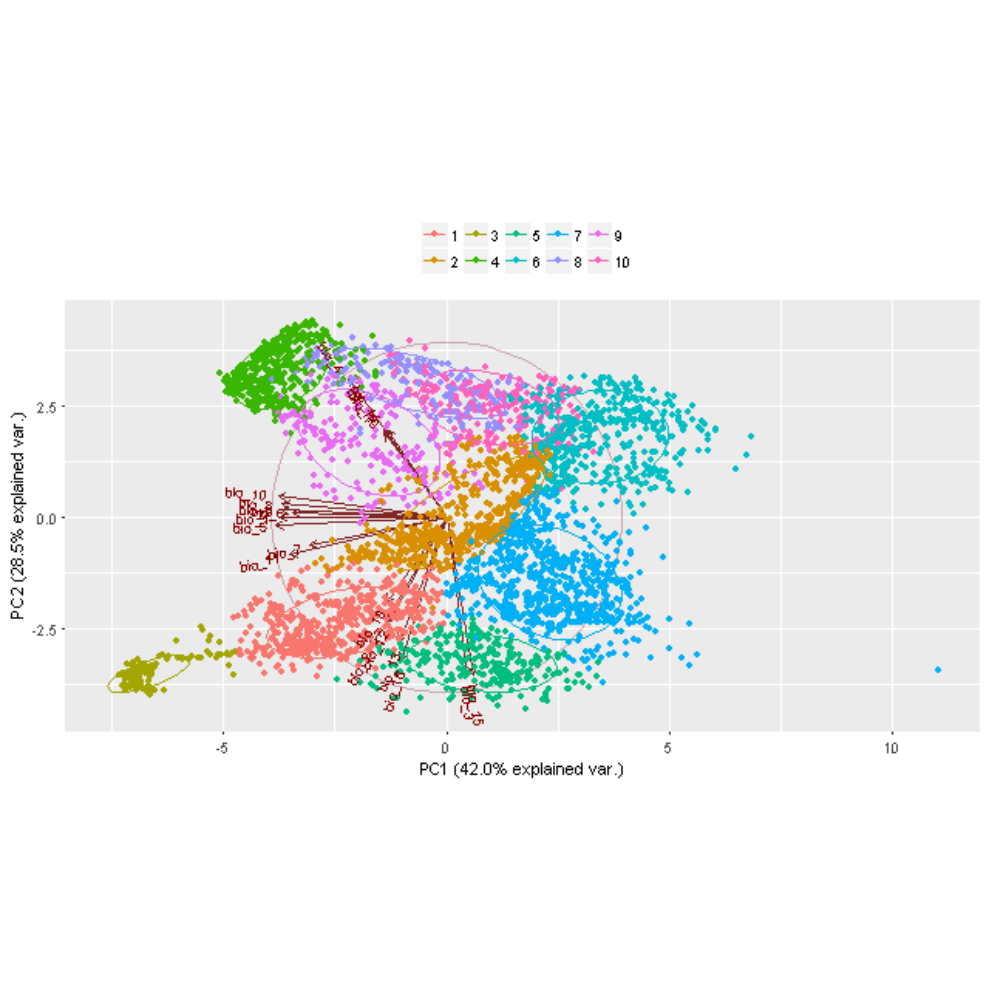

Figura 9: Espacio ambiental del café.

Caracterización de datos bioclimáticos

Espacio ambiental y teoría del nicho ecológico

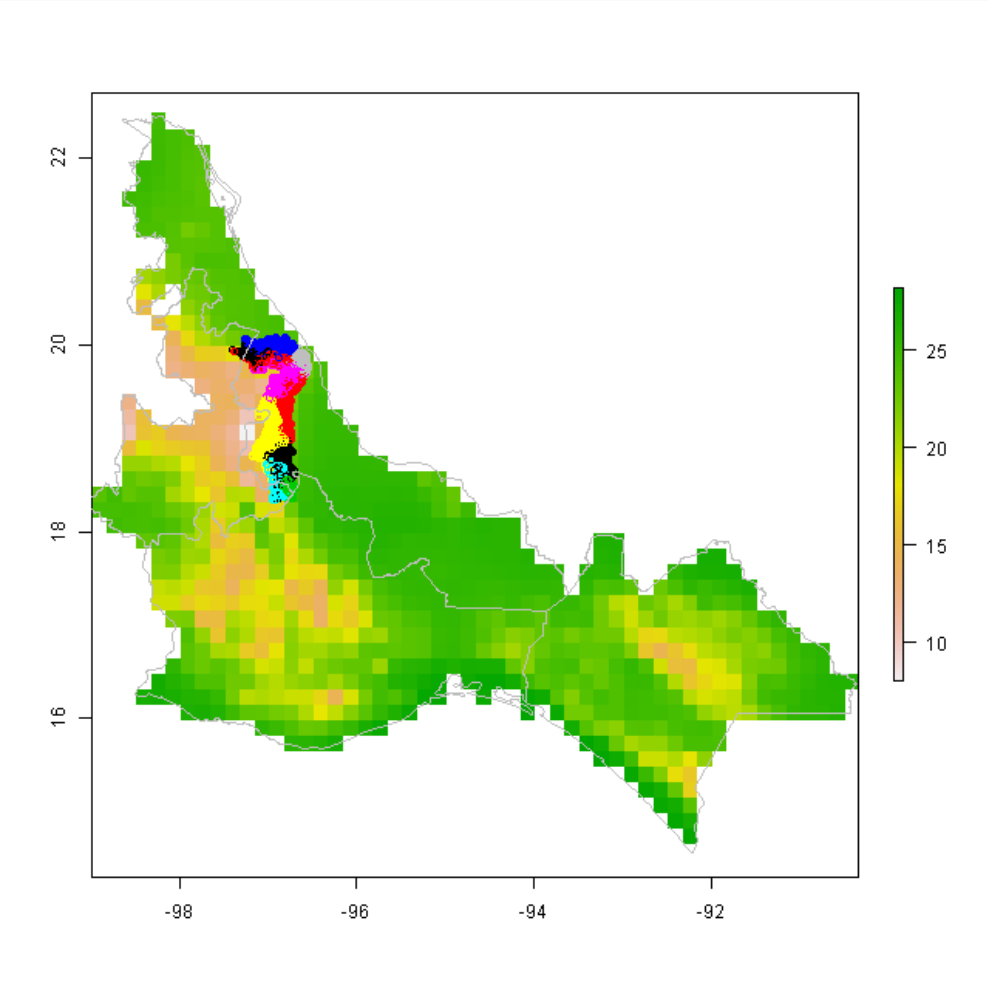

Figura 10: Espacio geográfico de las regiones de café.

Metodología del trabajo

Aprendizaje no supervisado (interpretable por un humano)

1. Identificar la región bioclimática correspondiente a cada plantación de café.
2. Identificar las características similares en las regiones bioclimáticas.
3. Establecer la relación entre calidad de café y región bioclimática.

Aprendizaje supervisado (no siempre interpretable)

1. Identificar la región bioclimática correspondiente a cada plantación de café.
2. Crear un vector que representa la información bioclimática de la plantación.
3. Ajustar un modelo predictivo para la calidad del café según parámetros bioclimáticos.

A pesar del parecido, ¡no son lo mismo!

Modelos de aprendizaje automático

Regresión lineal

Muy probablemente todos aquí han hecho regresión lineal alguna vez. Dado un conjunto de datos $\{y_i, x_{i1}, \ldots, x_{ip}\}_{i=1}^n$, un modelo de regresión lineal establece una relación lineal entre la variable dependiente y_i y el vector x_i dada por

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i, \quad i = 1, \ldots, n,$$

de manera que la relación entre la matriz de diseño

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}$$

y el vector de respuesta

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

Regresión lineal

Está dada por

$$\mathbf{y} = \mathbf{X}\beta + \epsilon,$$

donde

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

En palabras llanas esto significa que sin importar la complejidad o la cantidad de variables del problema, se asume que existe un hiperplano capaz de describir adecuadamente la relación entre todas las variables.

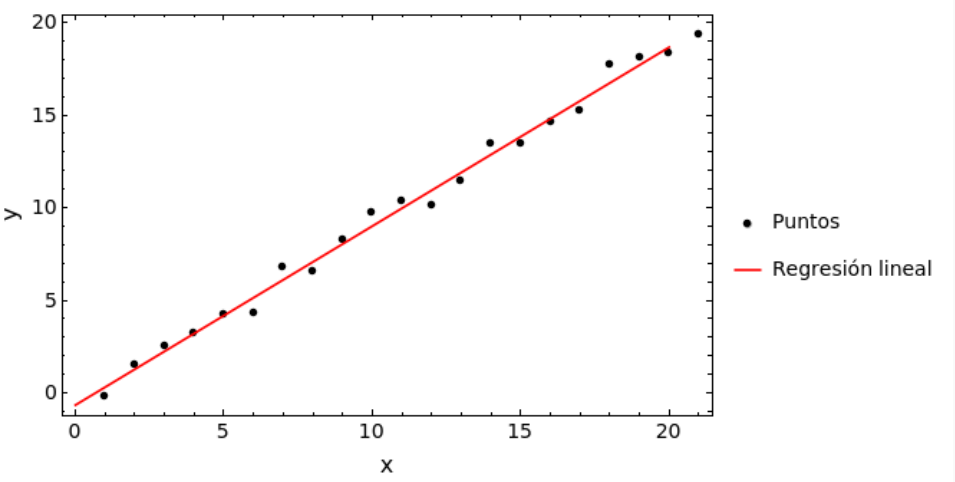

Figura 11: Regresión lineal cuando $n = 1$.

Regresión lineal

Ventajas	Desventajas
• Modelo extremadamente simple.	• No sirve para todos los conjuntos de datos.
• Por lo general da buenos resultados.	• No es bueno haciendo generalizaciones.
• Se puede conocer información acerca de lo que representa.	• En ocasiones requiere de ajustes “manuales”.

Redes neuronales artificiales

Es un modelo de ajuste basado (muy ligeramente) en el sistema nervioso. El modelo mapea desde un vector de entrada $\mathbf{x} \in \mathbb{R}^N$ a $f(\mathbf{x}) \in \mathbb{R}^M$, donde N y M son enteros positivos.

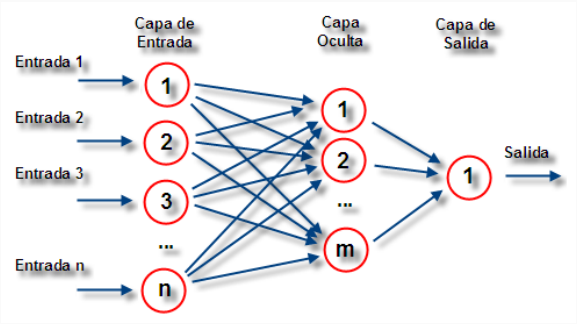

Figura 12: Diagrama de RNA.

Se define a f como

$$f_k(\mathbf{x}) = \sigma\left(\sum_i w_{ij}^k x_i^k + b_i^k\right),$$

donde σ es la función de activación, w_{ij} se denomina el *peso*, esto es la intensidad de conexión entre la neurona i y j, y b_i es el *bias* de la neurona i.

Redes neuronales artificiales

Por razones históricas es muy común el uso de la función sigmoide como función de activación, esto es

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

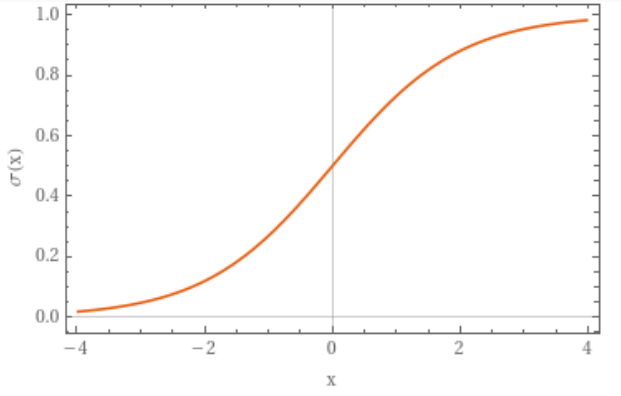

Figura 13: Función sigmoide.

El resultado de la propagación (*feedforward*) red neuronal artificial es

$$f(\mathbf{x}) = f_k \circ f_{k-1} \circ \cdots \circ f_1(\mathbf{x}).$$

Por eficiencia computacional suele ser mejor utilizar la notación matricial, de manera que el *feedforward* se describe como

$$\begin{aligned}
\mathbf{x}^2 &= \sigma\left(\mathbf{w}^1 \cdot \mathbf{x}^1 + \mathbf{b}^1\right), \\
\mathbf{x}^3 &= \sigma\left(\mathbf{w}^2 \cdot \mathbf{x}^2 + \mathbf{b}^2\right), \\
&\vdots \\
f(\mathbf{x}) = \mathbf{x}^D &= \sigma\left(\mathbf{w}^{D-1} \cdot \mathbf{x}^{D-1} + \mathbf{b}^{D-1}\right).
\end{aligned}$$

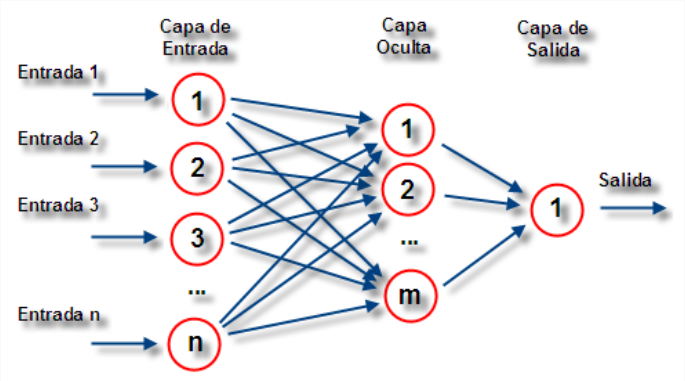

Figura 14: Diagrama de RNA.

Redes neuronales artificiales

Para ajustar el modelo es necesario variar los *pesos* y *bias* de forma que se aproximen al resultado correcto. A este proceso se le suele llamar *entrenamiento*. Para realizar el entrenamiento es necesario definir una función de costo, que mide la distancia entre el resultado de la RNA y el resultado esperado.

$$C(w, b) = \frac{1}{2n} \sum_i |y(x_i) - f(x_i)|^2$$

Durante el entrenamiento, por cada iteración se hacen variar los *pesos* y los *bias*.

$$w^k \to w_k - \eta \frac{\partial C}{\partial w^k}, \quad b^k \to b^k - \eta \frac{\partial C}{\partial b^k}.$$

El aplicar sucesivamente las transformaciones de w y b a capas anteriores se le conoce como retropropagación. ¿Por qué las RNA son tan buenas generalizando? Nadie lo tiene muy claro.

Redes neuronales artificiales

Ventajas	Desventajas
• Gran capacidad de generalización. • Se ha observado eficacia en gran cantidad de aplicaciones. Reconocimiento de imágenes, chatbots, análisis financiero, etc	• No existe interpretación directa de los valores aprendidos en w y b. • Requieren una muestra grande para garantizar un buen resultado. • El aprendizaje puede ser lento.

http://neuralnetworksanddeeplearning.com/

Resultados

Resultados predicción de calidad de café

OL_TOT a partir de datos bioclimáticos. Los resultados difieren.

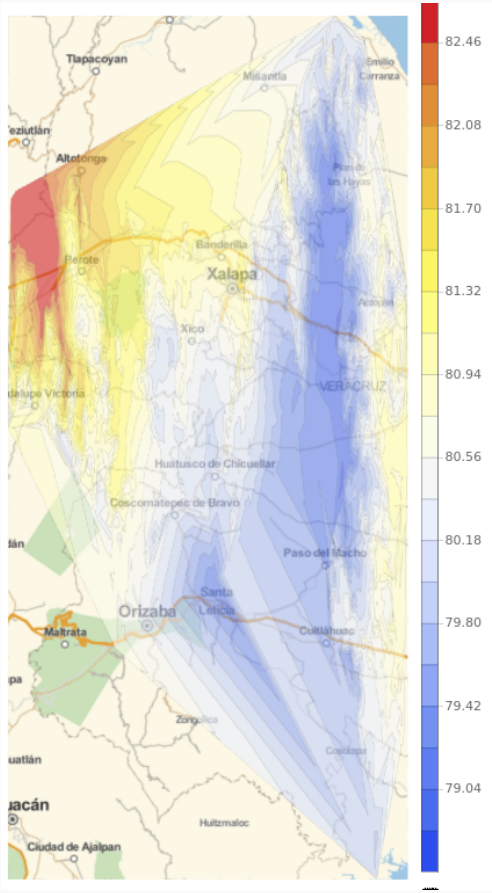

Figura 15: Usando regresión lineal.

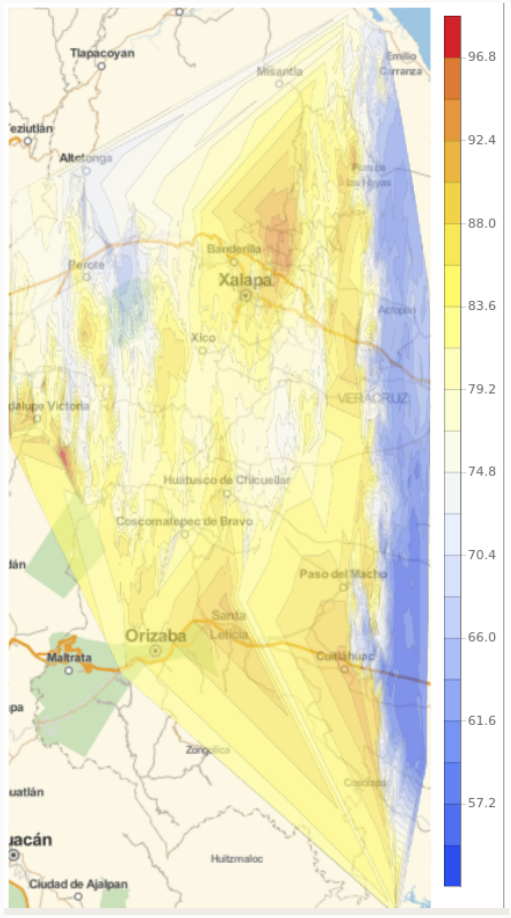

Figura 16: Usando redes neuronales artificiales.

Resultados predicción de calidad de café

Para ver qué modelo dio mejores resultados se usa la gráfica residual.

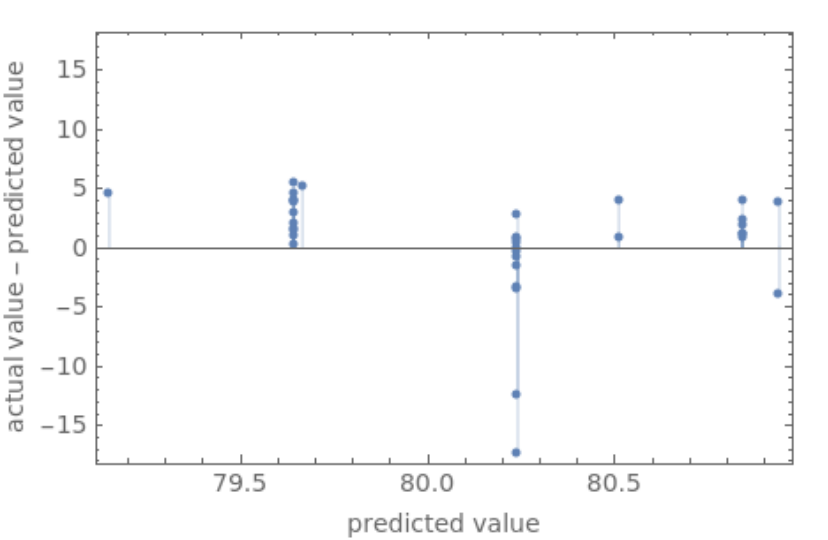

Figura 17: Usando regresión lineal. $\hat{\sigma} = 4{,}5644$.

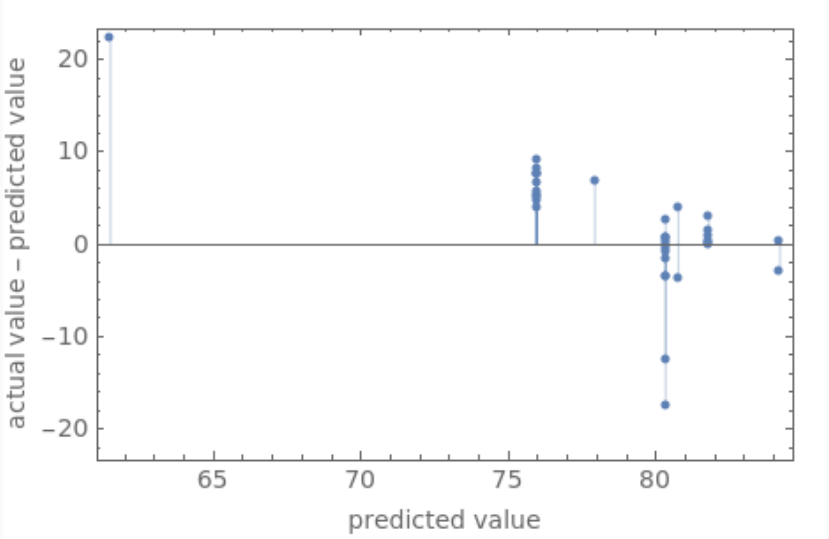

Figura 18: Usando redes neuronales artificiales. $\hat{\sigma} = 6{,}7248$.

Conclusión

- Se encontró una relación entre el índice de taza de excelencia y las variables ambientales
- El modelo permite predecir diferentes escenarios
- Es decir es posible cambiar escenarios climáticos (Clima de otros lugares o clíma del futuro)

- Es posible aplicar diferentes enfoques para resolver problemas emergentes
- Actualmente el país presenta problemas emergentes en áreas prioritarias
- Los problemas emergentes son multidisciplinarios
- Es un área de oportunidad para los físicos.