UNIVERSIDAD PERUANA UNION

FACULTAD DE INGENIERÍA Y ARQUITECTURA

Escuela Profesional de Ingeniería de Sistemas

Predicción del cambio climático con naive bayes.

Por:
Kenyi Simons López Azaña

Asesor:
Dr. Jorge Alejandro Sánchez Garcés

Juliaca, diciembre de 2019

DECLARACIÓN JURADA
DE AUTORÍA DEL TRABAJO DE
INVESTIGACIÓN

Dr. Jorge Alejandro Sánchez Garcés, de la Facultad de Ingeniería y Arquitectura, Escuela Profesional de Ingeniería de Sistemas, de la Universidad Peruana Unión.

DECLARO:

Que el presente trabajo de investigación titulado: "Predicción del cambio climático con Naive Bayes" constituye la memoria que presenta el estudiante Kenyi Simons López Azaña, para aspirar al grado de bachiller en Ingeniería de Sistemas, cuyo trabajo de investigación ha sido realizado en la Universidad Peruana Unión bajo mi dirección.

Las opiniones y declaraciones de este trabajo de investigación son de entera responsabilidad del autor, sin comprometer a la institución.

Y estando de acuerdo, firmo la presente constancia en Juliaca, a los 5 día del mes de diciembre del año 2019.

Dr. Jorge Alejandro Sánchez Garcés

Predicción del cambio climático con Naive Bayes

TRABAJO DE INVESTIGACIÓN

Presentada para poder optar el grado de bachiller de Ingeniería de Sistemas

JURADO CALIFICADOR

Mtro. Lennin Henry Centurión Julca
Presidente

Mg. Abel Angel Sullon Macalupu
Secretario

Ing. Angel Rosendo Condori Coaquira
Vocal

Dr. Jorge Alejandro Sánchez Garces
Asesor

Juliaca, 2 de diciembre de 2019

Predicción del cambio climático con Naive Bayes

López Azaña Kenyi Simons

EP. Ingeniería de Sistemas, Facultad de ingeniería y arquitectura de la Universidad Peruana Unión Filial Juliaca

Resumen

El análisis fue basado en una data set de 1800 registros que fueron clasificados con su variable temperatura y humedad donde 80% de la data fue utilizado para el entrenamiento y 20% para la prueba donde el error cuadrado para nuestro algoritmo fue de 0.22. También se utilizaron líneas temporales ARIMA para visualizar como seguirá la temperatura, la humedad y la precipitación en un futuro.

Palabras clave: Naive Bayes, Support vector machine, autoregressive integrated moving average, lineas temporales.

Abstract

The analysis was based on a data set of 1800 records that were classified with their temperature and humidity characteristics where 80% of the data was used for training and 20% for the test where the square error for our algorithm was 0.22. ARIMA time lines were also used to visualize how temperature, humidity and precipitation will follow in the future.

Keywords: Naive Bayes, Support vector machine, autoregressive integrated moving average, time lines.

1. Introducción

Naive Bayes, se conoce como un clasificador probabilístico simple que se basa en el teorema de Bayes que tiene suposiciones de independencia fuertes (ingenuas) entre las características. El proceso de clasificación bayesiano se define como un proceso que estima la probabilidad de una nueva observación perteneciente a una categoría predefinida, utilizando un modelo de probabilidad de acuerdo con la teoría de Bayes (P.~Cheeseman and J.~Stutz 1996). Naive Bayes estima la probabilidad previa de cada categoría en función de un gran conjunto de datos de entrenamiento, que se describen mediante una serie de variables, y supone que la clasificación podría estimarse calculando la función de densidad de probabilidad condicional y la probabilidad a posteriori (Ahmed et al. 2017).
ARIMA (autoregressive integrated moving average), son series de tiempo con modelos autor regresivos que funcionan con regresión y datos estadísticos con el fin de la predicción hacia el futuro (Fuente 2008).

Según la OMM (Organización Meteorológica Mundial o WMO siglas en inglés) menciona que las causas y efectos del cambio climático ha ido aumentando en vez de reducirse. También menciona que durante los periodos de 2015 al 2019 la emisión del CO2 fue casi al 20% superior a la de 5 años anteriores.
(Letcher 2018) menciona que el cambio climático se debe a causa del calentamiento global por la emisión del CO2 (dióxido de carbono un gas que es una de las causas del calentamiento global) fue aumentando un 0.3% en los últimos años. El cambio del clima, ha ido ocasionando desastres por diferentes lugares cuyos desastres son: la lluvia intensa, tormentas eléctricas entre otros.

(Shashaani et al. 2018) menciona que las tormentas son un peligro natural prominente que ocasiona pérdidas materiales, esto descrito por la cantidad de humedad, precipitación y temperatura. En el Norte del Perú se encuentran lugares donde el clima y el lugar en el que se encuentran son más susceptibles a sufrir inundaciones o huaicos a causa de las lluvias. (INGEMMET 2017).
Por otra parte, (Indeci 2012; Senamhi 2019), corroboran que en las últimas décadas se registraron muchas muchos desastres naturales como: 44 huaycos, 140 deslizamientos, 104 derrumbes, 258 inundaciones y 1463 lluvias intensas, donde 82.86% de viviendas han sido afectadas y 17.14% de viviendas fueron destruidas. Sin mencionar la tasa de mortalidad por dichos desastres. Y las temperaturas mínimas han ido disminuido extremadamente en un 10% más abajo. En este sentido el cambio climático puede afectar de manera peligrosa a las comunidades.

Los climas al no ser pronosticables ocasionan pérdidas materiales como también pérdidas personales sin poder hacer las prevenciones respectivas. Las aplicaciones y métodos existentes como NWP (Numeral Weather Prediction), MRF (Medium Range Prediction), han permitido las predicciones de fenómenos climatológicos. Sin embargo, los modelos de predicción no son perfectos por lo que la precisión no es tan exacta (Lorenc 1986). Esto proporciona la oportunidad de aplicar modelos machine learning para la predicción del clima y ver la precisión de estos modelos. Ya que se siguen utilizando los modelos tradicionales.

La predicción del clima juega un papel vital en estos tiempos, los sectores agrícolas, son los que más dependen de esto. Un trabajo realizado por Fente (2018) quien realizo una predicción del tiempo utilizando redes neuronales, donde se manejaron teorías como las redes neuronales recurrentes, cómo funcionan y para qué sirven, donde las variables tomadas por el investigador fueron la precipitación, la temperatura la humedad, presión, punto de roció y la visibilidad.

Otro Segundo trabajo desarrollado por Das (2018) que se denomina “Pronóstico meteorológico basado en estimador bayesiano utilizando WSN” donde utilizo herramientas de machine learning, para especificar el autor de este trabajo utilizo los modelos de Naive Bayes, Svm. Tomando como variables la Temperatura, Temperatura aparente, Humedad, Velocidad del viento, visibilidad y presión. Todo para poder mejorar el proceso de estimación y predicción. Este estudio demostró la efectividad de naive bayes, y la precisión de datos obtenidos de los sensores.

Otro tercer trabajo desarrollado por Haupt et al.(2018) que se denomina aprendizaje automático aplicada a la predicción climática utiliza herramientas que hacen uso de machine learning ADIcast esta herramienta usa métodos como el NWP(numeral weather prediction), donde usa variable de temperatura, punto rocío, precipitación. Este trabajo tuvo como finalidad solo de demostrar como la herramienta se utiliza para la predicción.

Todos estos trabajos se relacionan con la investigación en curso que toma las mismas variables añadiendo otras variables como la velocidad del viento, la dirección del viento, las temperaturas mínimas, máximas y promedios, y se propone utilizar algoritmos bayesianos para la predicción del clima. Por lo tanto el objetivo de esta investigación es utilizar el modelo naive bayes y adicionalmente utilizar modelos ARIMA, para la predicción del clima y para poder detectar patrones del clima. Las variables que se incluyen aparte de las otras investigaciones vistas son las temperaturas mínimas, máximas, la dirección del viento y la velocidad del ciento

2. Materiales y Métodos

2.1. Participantes

Arquitectura: en esta investigación realizada se propuso una arquitectura de solución *Figura 1* donde se utilizó una data de validación y data de entrenamiento para luego usar los algoritmos de naive bayes y support vector machine y mostrar los resultados y las precisiones de cada modelo, se utilizó el svm solo para ver que el algoritmo que seleccionamos es mejor que otro de clasificación.

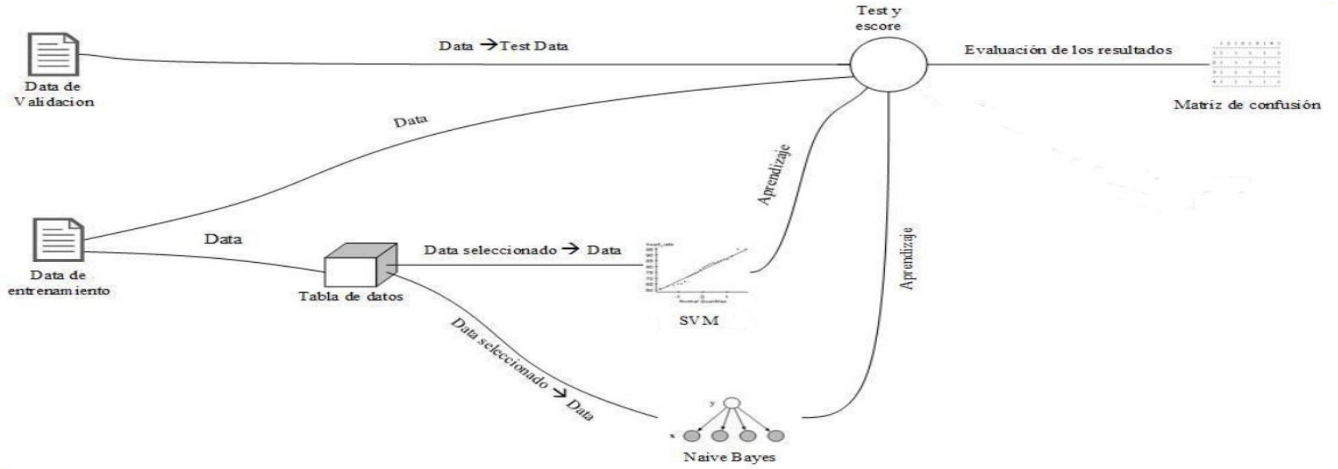

Figura 1. arquitectura de solución aplicando naive bayes y support vector machine para lograr comparar la precisión entre esos dos modelos.

2.2. Instrumentos

Tabla 1

Materiales utilizados para el desarrollo de la investigación

Materiales	versión	licencia	descripción
Azure		Libre	un servicio gratuito para desarrollar y ejecutar código directamente en un navegador usando lenguajes como Python 2, Python 3, R, F# y los paquetes de librerías más populares como Anaconda, el cual por cierto ya viene preinstalado (¿Qué es Azure Notebooks? – Azurebrains n.d.). El código desarrollado en esta investigación fue hecho en azure notebooks.
Python	3.6	Libre	es un lenguaje de programación de alto nivel, cuya filosofía hace hincapié en una sintaxis que favorezca un código legible. Es un lenguaje multiparadigma porque soporta programación orientada a objetos como también programación funcional que fue creado por Python Software Fundation.
Notebooks		Libre	es un cuaderno para poder programar en los lenguajes python, R. viene bibliotecas preinstaladas para hacer machine learning, ciencia de datos, redes neuronales entre otras tareas.
csv		Libre	El contenido almacenado en el formato CSV se refieren a archivos de datos adjuntos con él .csv extensión, y estos archivos CSV también se llaman valores separados por comas archivos. El "CSV" en un archivo de puesta con él .csv extensión significa "valores separados por comas" debido a que los datos de estos archivos CSV son detalles dividido por comas en conjuntos particulares de información.

2.3. Metodología

La data utilizada para esta investigación fue sacada de manera gratuita de la senamhi con 1800 datos, estos fueron que fueron tomados por día desde el 2014 hasta inicios del 2019:

Tabla 2

Datos sacados de la senamhi.

	temperatura_prom	temperatura_max	temperatura_min	humedad	precipitacion	presion	velocidadV	direccionV
0	9.91	17.7	0.5	36.13	0	644.21	1.34	277
1	7.88	16.7	-2.6	44.13	0	645.23	1.68	285
2	8.89	17.9	-0.2	57.54	0	656.37	1.73	280
3	9.69	17.4	2.6	49.71	0	646.76	2.09	312
4	10.17	17.8	3.2	40	0	645.92	2.01	281

En la data sacada de la senamhi graficada para hallar el nivel de dispersión de los datos *Figura 2.*, la *Figura 3* muestra los datos que se correlacionan de las variables tomadas por la data set, el valor va de -1 a +1 donde las variables se correlacionan, un valor de 0 significa que no hay relación entre dos variables sin embargo si es diferente de 0 hay relación entre las variables, si es para el negativo las relaciones es negativa.

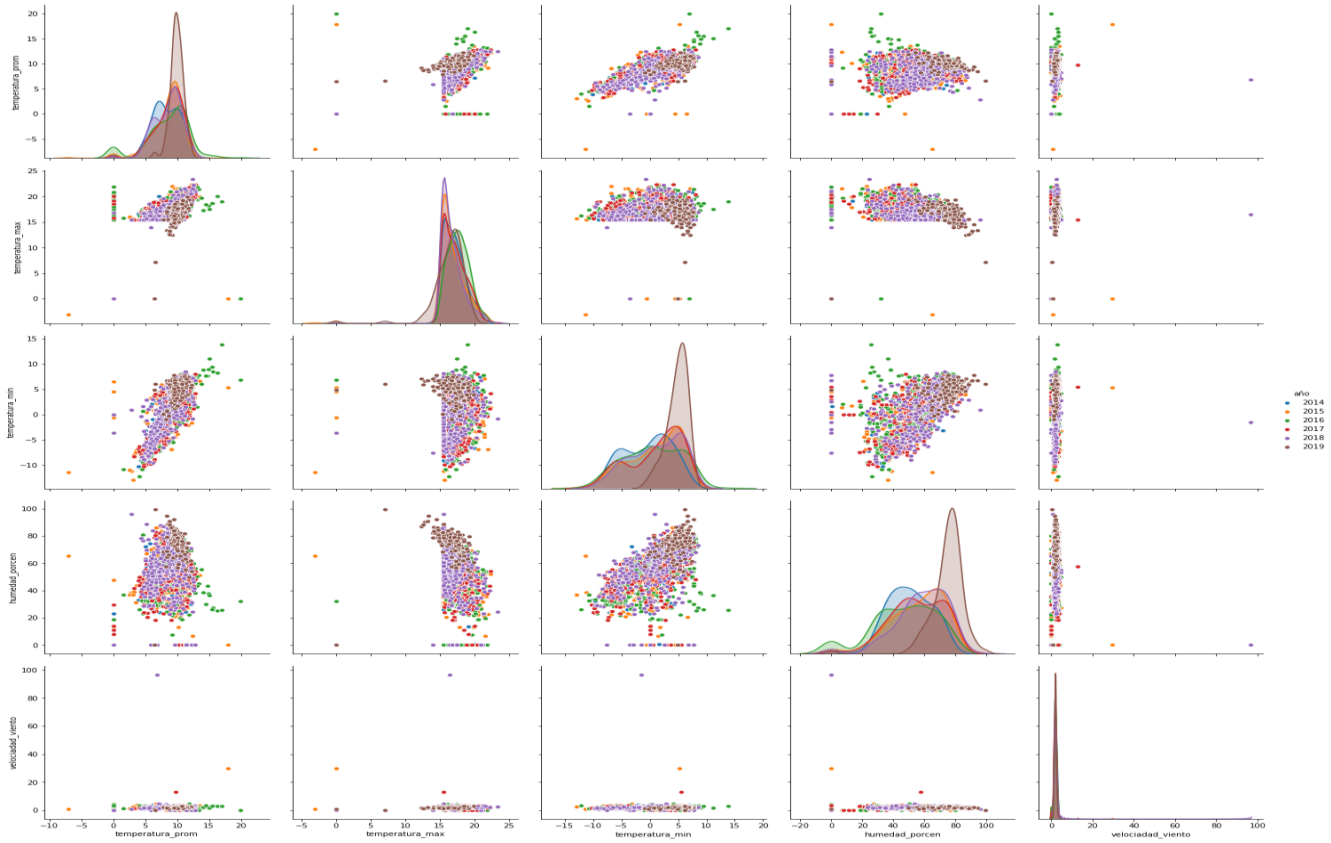

Figura 2. Dispersión de los datos, donde se observa que la data tiene muchos datos dispersos

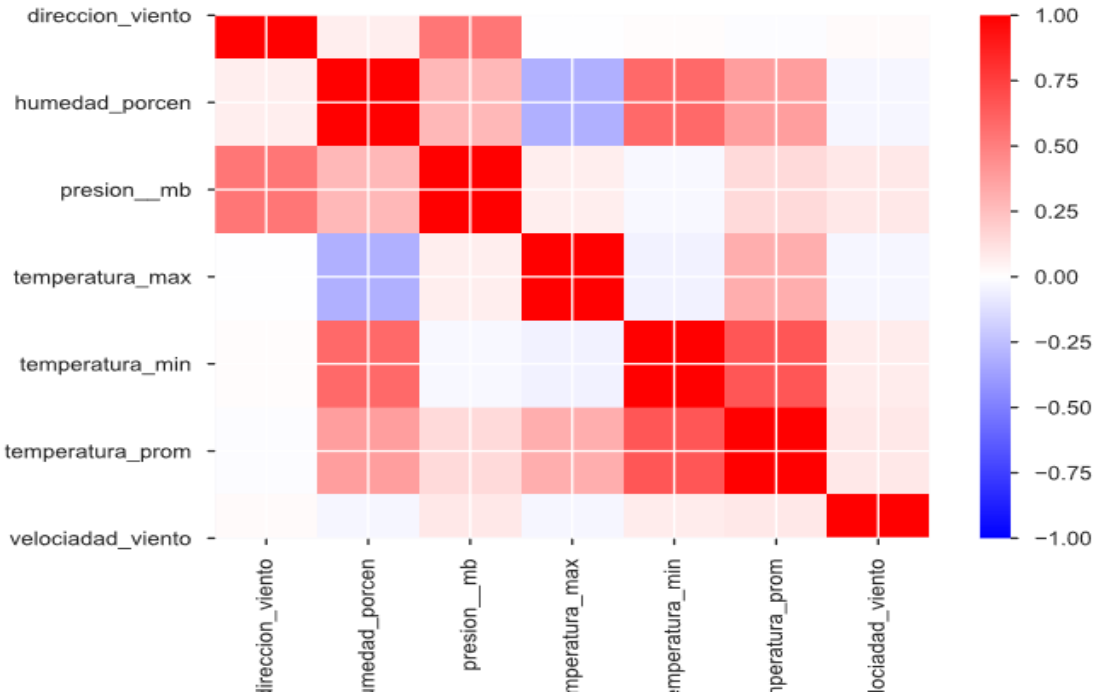

Figura 3. Correlación de pearson del dataset donde se ve los features más relevantes.

Para la preparación de los datos de 1800 registros se limpió los datos, features que no tenían ni un valor, luego se separó la data en una de entrenamiento y otra de testing para poder utilizar el modelo. Los modelos utilizados dieron el svm y el naive bayes (gaussian).

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

Fórmula 1. Los parámetros σ_y y μ_y son estimados usando la máxima predicción.

Donde obtuvimos dos resultados de ambos algoritmos uno es el que se propuso en esta investigación, y el otro es el que se propuso para validar que el modelo que nosotros elegimos es el más óptimo.
En la *tabla 3* podemos observarla precisión que nos mostró el modelo propuesto por nosotros con un 87% de precisión y su error cuadrado de 43% mientras que el algoritmo de svm nos mostró una precisión del 67% de precisión. Dando a reconocer que nuestro algoritmo propuesto es mejor que el otro algoritmo.

Tabla 3

Porcentaje de precisión entre naive bayes y svm.

Modelos	***Precisión***	***Error cuadrado***
Naive Bayes (gaussian)	0.8887328767123287	0.4328767123287671
SVM	0.6794520547945205	0.8383561643835616

Para validar que nuestras variables que tomamos tienen una influencia en el entrenamiento sacamos la temperatura máxima y mínima para ver el accuracy y el resultado boto de 87.60% de precisión. Mientras que al eliminar solo una de las variables que es la temperatura minina nos dio un resultado de 89.04%; después se eliminó una de las variables que es la temperatura máxima y se agregó la temperatura mínima dándonos un resultado del 87.80%; después colocamos las dos variables dándonos un resultado de 88.87%.

lo cual la variable que más influye es la temperatura máxima, mientras que la temperatura mínima no influye mucho *Tabla 4.*

Tabla 4

Porcentaje de precisión de las variables que se tomó para esta investigación

	variables	precision
Naive bayes	Sin temperatura máxima y mínima	87.60%
	Con temperatura máxima	89.04%
	Con temperatura mínima	87.70%
	Con temperatura máxima y mínima	88.87%

3. Resultados y Discusión

3.1. Resultados 1

En esta presente investigación el clasificador naive bayes fue utilizado para poder clasificar y predecir con la data set de la senamhi; las predicciones que se muestran son de días en el futuro, días que no se encuentran en la data set; dando valores que podría ocurrir. Donde el nivel de precisión del algoritmo seleccionado para esta investigación mostro una predicción del 87% que fue el máximo *Tabla 7.*

Tabla 6

Tabla de datos sacada de la senami desde el 19 de abril del 2019 para validar los datos predichos.

Dater/dias	Temp Promedio	Temp Máxima	Temp Mínima	Humedad	Precipitación	Presión	Velocidad del viento	Direccion del viento
19/04/2019	9.93	17.4	3.2	78.54	2.2	647	2.07	281
20/04/2019	10.1	17.6	5.3	78.58	0.6	646.7	1.12	232
21/04/2019	8.39	17.1	0.7	79.25	0.1	645.43	1.72	90
22/04/2019	9.43	16.2	5.1	83.79	7.9	645.1	1.61	254
23/04/2019	6.42	15.2	4.9	84.32	0.2	646.47	0.82	176
24/04/2019	10.69	18.5	6.1	76.17	2.3	646.67	2.86	7
25/04/2019	6.6	7.1	6.1	99.29	0	648.04	0.24	305

Tabla 7

Tabal de datos de la predicción, desde el 19 de abril del 2019; la data de entrenamiento solo era hasta el 18 de abril del 2019.

Dater/dias	Temp Promedio	Temp Máxima	Temp Mínima	Humedad	Precipitación	Presión	Velocidad del viento	Direccion del viento	Naive Byes
19/04/2019	9.2203	17.7121	2.1098	68.5517	5.0933	644.7124	1.4334	29.2260	0.87
20/04/2019	7.3855	16.4173	-1.3760	68.0651	2.6185	644.6254	1.7544	211.1007	0.8
21/04/2019	8.8277	17.7987	1.7901	73.2944	0.9543	644.4675	1.8897	19.1467	0.75
22/04/2019	9.4904	16.3963	6.0280	81.5362	-0.1483	645.2118	2.3035	38.3647	0.85
23/04/2019	8.0218	15.5097	2.6484	87.6252	14.6321	645.6441	2.1660	234.4402	0.83
24/04/2019	9.2371	15.4773	6.2694	82.4075	0.0678	645.7903	1.4492	-29.7436	0.7
25/04/2019	8.0589	15.5280	3.6603	84.3106	5.2746	646.3124	2.7671	224.7935	0.86

4. Conclusiones

El presente estudio proporciona un modelo predictivo que utiliza un modelo de Naive bayes para poder clasificar y predecir los cambios meteorológicos. Pero estas predicciones no son del todo precisas, porque con la contaminación que hoy en día se vive el clima está cambiando drásticamente, predecir algún tipo de clima de los lugares es un poco complicado; en un lugar puede estar lloviendo mientras en otro está soleando. La data sacada en esta investigación es de un solo lugar que es el centro de la cuidad de Juliaca, pero hay lugares aledaños a la ciudad que es muy diferente el comportamiento del clima, mientras en el centro llueve a los alrededores está soleando. Una solución sería que en cada lugar se pondríamos más sensores meteorológicos para tener una mejor precisión en la predicción del clima.

Para una investigación que se realizara en el futuro se podrá pronosticar desastres naturales como los huaicos u otro desastre siguiendo estos patrones meteorológicos.

Referencias

"¿Qué Es Azure Notebooks? – Azurebrains." https://www.azurebrains.com/2019/06/14/que-es-azure-notebooks/ (October 30, 2019).

Ahmed, Md Shakil, Md Shahjaman, Md Masud Rana, and Md Nurul Haque Mollah. 2017. "Robustification of Naïve Bayes Classifier and Its Application for Microarray Gene Expression Data Analysis." *BioMed Research International* 2017.

Das, Rik. 2018. "Bayesian Estimator Based Weather Forecasting Using WSN." *2018 3rd International Conference and Workshops on Recent Advances and Innovations in Engineering (ICRAIE)* 2018(November): 1–4.

Fente, Dires Negash. 2018. "Weather Forecasting Using Artificial Neural Network." *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)* (Icicct): 1757–61.

Fuente, Santiago. 2008. "Modelo Arima." *Universidad Autonoma de Madrid*: 100.

Haupt, Sue Ellen et al. 2018. "Machine Learning for Applied Weather Prediction." *Proceedings - IEEE 14th International Conference on eScience, e-Science 2018*: 276–77.

Indeci, D E L. 2012. "Compendio Estadístico Del Indeci 2011."

INGEMMET. 2017. "Inundaciones y huaicos en la costa norte: Causas y efectos - Noticias - Ingemmet." http://www.ingemmet.gob.pe/-/inundaciones-y-huaicos-en-la-costa-norte-causas-y-efectos (June 30, 2019).

Letcher, Trevor M. 2018. Managing Global Warming *Why Do We Have Global Warming?* Elsevier Inc. http://dx.doi.org/10.1016/B978-0-12-814104-5.00001-6.

Lorenc, A. C. 1986. "Analysis Methods for Numerical Weather Prediction." *Quarterly Journal of the Royal Meteorological Society* 112(474): 1177–94.

Mcquade, Scott. 2013. "Climate Informatics : Accelerating." *Computing in Science & Engineering*: 32–40.

P.~Cheeseman, and J.~Stutz. 1996. "Bayesian Classification ({AutoClass}): Theory and Results." *Advances in Knowledge Dicovery and Data Mining*: 153–80.

Senamhi. 2019. "SENAMHI - Perú." https://www.senamhi.gob.pe/?p=cambio-climatico (June 24, 2019).

Shashaani, Sara et al. 2018. "Multi-Stage Prediction for Zero-Inflated Hurricane Induced Power Outages." *IEEE Access* 6: 62432–49.

"scikit-learn: machine learning in Python — scikit-learn 0.21.3 documentation." [Online]. Available: https://scikit-learn.org/stable/. [Accessed: 11-Nov-2019].

"¿Qué Es Azure Notebooks? – Azurebrains." https://www.azurebrains.com/2019/06/14/que-es-azure-notebooks/ (October 30, 2019).

Ahmed, Md Shakil, Md Shahjaman, Md Masud Rana, and Md Nurul Haque Mollah. 2017. "Robustification of Naïve Bayes Classifier and Its Application for Microarray Gene Expression Data Analysis." *BioMed Research International* 2017.

Das, Rik. 2018. "Bayesian Estimator Based Weather Forecasting Using WSN." *2018 3rd International Conference and Workshops on Recent Advances and Innovations in Engineering (ICRAIE)* 2018(November): 1–4.

Fente, Dires Negash. 2018. "Weather Forecasting Using Artificial Neural Network." *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)* (Icicct): 1757–61.

Fuente, Santiago. 2008. "Modelo Arima." *Universidad Autonoma de Madrid*: 100.

Haupt, Sue Ellen et al. 2018. "Machine Learning for Applied Weather Prediction." *Proceedings - IEEE 14th International Conference on eScience, e-Science 2018*: 276–77.

Indeci, D E L. 2012. "Compendio Estadístico Del Indeci 2011."

INGEMMET. 2017. "INUNDACIONES Y HUAICOS EN LA COSTA NORTE: CAUSAS Y EFECTOS - Noticias - Ingemmet." http://www.ingemmet.gob.pe/-/inundaciones-y-huaicos-en-la-costa-norte-causas-y-efectos (June 30, 2019).

Letcher, Trevor M. 2018. Managing Global Warming *Why Do We Have Global Warming?* Elsevier Inc. http://dx.doi.org/10.1016/B978-0-12-814104-5.00001-6.

Lorenc, A. C. 1986. "Analysis Methods for Numerical Weather Prediction." *Quarterly Journal of the Royal Meteorological Society* 112(474): 1177–94.

Mcquade, Scott. 2013. "Climate Informatics : Accelerating." *Computing in Science & Engineering*: 32–40.

P.~Cheeseman, and J.~Stutz. 1996. "Bayesian Classification ({AutoClass}): Theory and Results." *Advances in Knowledge Dicovery and Data Mining*: 153–80.

Senamhi. 2019. "SENAMHI - Perú." https://www.senamhi.gob.pe/?p=cambio-climatico (June 24, 2019).

Shashaani, Sara et al. 2018. "Multi-Stage Prediction for Zero-Inflated Hurricane Induced Power Outages." *IEEE Access* 6: 62432–49.