

PREDICCIÓN Y VISUALIZACIÓN DE DATOS DEL CRIMEN

PREDICTION AND VISUALIZATION OF CRIME DATA

Huancapaza Hilasaca, Liz Maribel¹; Vargas Belizario, Ivar²;
Arpasi Chura, Rodolfo Fredy³

Resumen

La predicción y visualización de datos del crimen, actualmente tienen una notable importancia porque pueden proporcionar información relevante cómo, dónde y cuándo puede suceder un determinado tipo de crimen. Esta información puede ayudar a crear planes anticipados de acción contra el crimen. En ciencia de la computación desde el enfoque de aprendizaje de máquina, la predicción es realizada por medio de propuestas de algoritmos de clasificación supervisada. Por otro lado, la visualización de datos del crimen es aplicada por medio de gráficos, de mapas de calor que ayudan al monitoreo de la actividad criminal. En este trabajo proponemos un método para predecir y visualizar datos reales del crimen. El método propuesto está basado en la transformación del conjunto de datos por medio de la creación de nuevos atributos a partir de atributos existentes. Donde el usuario interactúa mediante la visualización del conjunto de datos para mejorar la calidad de la clasificación. Los resultados iniciales indican una calidad de clasificación alcanzada del 91,00 %. Tareas de interacción del usuario en el proceso de creación de nuevos atributos puede ayudar a mejorar los resultados de clasificación de datos del crimen.

Palabras clave: Predicción, clasificación, visualización, datos del crimen.

-
- 1 Estudiante Especial de Posgraduación en Ciencia de la Computación, ICMC, *Universidade de São Paulo*, Brasil.
 - 2 Doctorando en Ciencia de la Computación, ICMC, *Universidade de São Paulo*, Brasil.
 - 3 Profesor de la Escuela Profesional de Ingeniería de Sistemas, Universidad Andina Néstor Cáceres Velásquez,

Abstract

The prediction and visualization of crime data, currently have a remarkable importance because they can provide relevant information about how, where and when a certain type of crime can happen. This information can help create anticipated action plans against crime. In computer science, from the machine learning approach, the prediction is carried out through proposals of supervised classification algorithms. On the other side, the visualization of crime data is applied by means of heat map graphics that help to monitor criminal activity. In this work we propose a method to predict and visualize real crime data. The proposed method is based on the transformation of the data set through the creation of new attributes from existing attributes. Where the user interacts by the displaying the visualization of data set to improve the quality of the classification. The initial results indicate a classification quality reached of 91,00%. User interaction tasks in the process of creating new attributes can help improve the results of classification of crime data.

Keyword: *Prediction, classification, visualization, crime data.*

Introducción

El crimen es un problema social muy complejo de resolver, es así que en ciudades con alto índice de criminalidad es necesario utilizar herramientas o métodos computacionales para predecir el crimen con un alto grado de precisión. Todos los días, datos de crimen son generados en grandes cantidades y a la vez son almacenados por los departamentos policiales. Estos datos almacenan información muy valiosa y es necesario analizarlos para extraer patrones o información relevante como por ejemplo dónde y cuándo sucederá un determinado tipo de crimen (predicción del crimen). De esta forma la predicción del crimen puede proporcionar información relevante para ayudar a crear planes anticipados para combatir la criminalidad, en este contexto es importante y necesario desarrollar nuevas herramientas o métodos computacionales para mejorar la calidad de predicción del crimen. Actualmente, la creación de nuevas herramientas o métodos computacionales están siendo abordados por un nuevo campo científico que conjuga varias áreas de ciencia de la computación, estas áreas son: la visualización y análisis de datos, ciencia de datos y aprendizaje de máquina. Es en este nuevo campo científico donde finalmente el raciocinio analítico se desenvuelve por medio de interfaces visuales e interactivas.

Recientemente, métodos para predecir el crimen son formulados considerando el preprocesamiento de los datos y el entrenamiento supervisado de clasificadores, algunos de estos clasificadores son *Super Vector Machine*, *Naive Bayes*, *Random Forest*, entre otros (Sathyadevan *et al.*, 2014; Vineeth *et al.*, 2016; Shamsuddin *et al.*, 2017). Por otra parte los trabajos que incluyen la visualización de datos del crimen están orientados a visualizar las instancias de los crímenes por ejemplo en mapas geográficos (Jayaweera *et al.*, 2015; Wang, 2015). También *Heatmap*, es otro método de visualización frecuentemente utilizado, este método de visualización está basado en función a la densidad o incidencia de los crímenes, donde gráficamente es generado un mapa de calor basado en colores para representar la alta o baja incidencia (Wu *et al.*, 2017; Chae *et al.*, 2015).

En áreas de visualización, aprendizaje máquina y ciencia de datos, recientemente existe un nuevo enfoque para abordar problemas de clasificación de datos, este nuevo enfoque es conocido como Exploración Visual de Datos propuesto inicialmente para imágenes (Brandoli *et al.*, 2010). La exploración visual de datos tiene como propósito incluir al usuario de forma interactiva en las tareas que impliquen la clasificación de datos, esto con el objetivo de mejorar la calidad de clasificación.

Basado en el nuevo enfoque de exploración visual de datos, en este artículo proponemos un método para predecir datos del crimen teniendo como objetivo principal crear nuevos atributos a partir de los atributos existentes, esto es conocido como *Feature Engineering* (Fatemeh Nargesian, 2017). En nuestro método la predicción del crimen es realizada por algoritmos de clasificación supervisada (clasificadores). Y con el objetivo de mejorar la calidad de clasificación es empleado un método de visualización conocido como Coordenadas Paralelas (Gauthier-Villars, 1885). Los resultados iniciales de clasificación alcanzados con el método propuesto fueron de 91,00 % para 30 categorías de crimen, donde fueron creados 4 nuevos atributos.

Material y métodos

El conjunto de datos de crímenes empleado en la investigación fue: *San Francisco Crime Classification* (<https://www.kaggle.com/c/sf-crime>). El conjunto de datos de entrenamiento posee 878 049 registros de crímenes reales ocurridos desde el 1/1/2003 hasta el 5/13/2015 y que corresponde a 39

categorías desbalanceadas. Los atributos de este conjunto de datos son descritos a continuación:

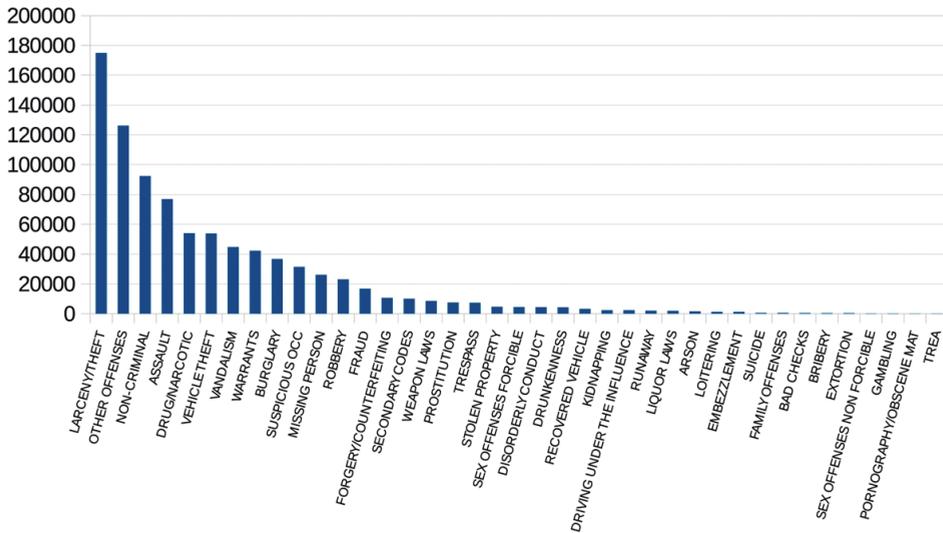
- *Dates*: fecha en formato dato tipo *time*.
- *DayOfWeek*: día de la semana.
- *PdDistrict*: identificador del departamento policial.
- *Address*: la dirección aproximada donde ocurrió el crimen.
- *X*: latitud.
- *Y*: longitud.
- *Category*: categoría del crimen (este es atributo a predecir, existen 39 categorías).
- *Description*: una nota que describe detalles pertinentes del crimen.
- *Resolution*: para indicar el estado del crimen; si fue resuelto o no.

Un ejemplo de 4 registros o instancias del conjunto de datos que fue empleado en el trabajo es ilustrado en la Tabla (1), por otra parte, la Figura (1) muestra un histograma resumiendo la cantidad de instancias por categoría de todo el conjunto de datos.

Tabla 1. Ejemplo de 4 instancias del conjunto de datos
San Francisco Crime Classification

Dates	DayOfWeek	PdDistrict	Address	X	Y	Category
2015-05-13 23:53:00	Wednesday	NORTHERN	OAK ST / LAGUNA ST	-122.4258	37.7745	WARRANT ARREST
2015-05-13 23:00:00	Wednesday	CENTRAL	24TH ST / CAPP ST	-122.4173	37.7523	ROBBERY
2015-02-21 16:00:00	Friday	TARAVAL	200 Block of 9TH ST	-122.4698	37.7170	LOST PROPERTY
2014-12-12 14:00:00	Friday	NORTHERN	900 Block of HYDE ST	-122.4170	37.7895	BATTERY

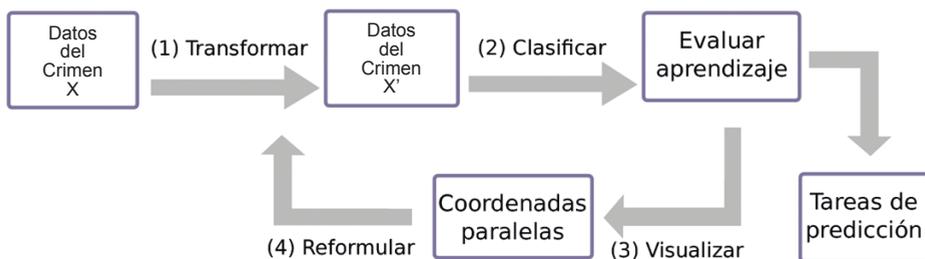
Figura 1. Histograma de las categorías del conjunto de datos *San Francisco Crime Classification*.



Propuesta

De forma similar que el método de exploración visual de datos formulado para imágenes (Brandoli *et al.*, 2010), proponemos un método para explorar visualmente datos del crimen con el objetivo de mejorar la calidad de clasificación y por ende la calidad de predicción. De esta forma la Figura (2) ilustra la metodología adoptada para el método de predicción de datos del crimen formulado. Es en el Algoritmo (1) donde se resume el método, siendo sus etapas explicadas en detalle a continuación:

Figura 2. Flujograma del método propuesto.



1. Transformar: en esta etapa son procesadas las tareas de preprocesamiento y la creación de nuevos atributos.

- a) Preprocesamiento: son procesados tareas para limpiar, completar y convertir datos. De esta forma los datos categóricos fueron convertidos a datos numéricos. Los atributos empleados del conjunto original fueron: F1: (X) latitud, F2: (Y) longitud, F3: día de la semana, F4: departamento policial y F5: mes.
- b) Creación de nuevos atributos: fueron creados 4 atributos, los cuales son descritos a continuación:
 - 1) *F6: días del mes, extraídos del atributo fecha tipo tiempo, los valores están en el rango [0-30].
 - 2) *F7: distancia, esta distancia es calculada entre la hora del crimen respecto a una hora determinada; este nuevo atributo es calculado según la formulación de la Ecuación (1).

$$d = |h - hd| \quad \text{Ecuación (1)}$$

donde, h es la hora actual y hd es una hora determinada, en este caso fue establecida en 12 horas, para evaluar la distancia en horas respecto al medio día.

- 4) *F8: hora, fue extraída del atributo fecha tipo tiempo, los valores están en el rango [0-23].
- 5) *F9: minutos, fue extraída del atributo fecha tipo tiempo, los valores están en el rango [0-59].

Después el conjunto de datos es balanceado por clase, para balancear el conjunto de datos fue empleado la replicación de instancias en las clases minoritarias, el método empleado fue *oversampling* (Gutierrez, 2015).

2. Clasificar: en esta etapa es entrenado un clasificador de forma supervisada, para este caso fue empleado el clasificador *Random Forest*. Una vez entrenado el clasificador es evaluado la calidad de clasificación. Si la calidad de clasificación es adecuada, entonces el clasificador ya puede ser empleado para tareas de predicción del crimen. En el caso de que no haya sido alcanzada una adecuada calidad de clasificación,

se procede a visualizar el conjunto de datos en la etapa siguiente (3). Para evaluar la calidad de clasificación fue empleada la métrica de *Accuracy* (Marsland, 2014). Donde dada una matriz de confusión M_{ij} que es el producto del resultado de la clasificación, la métrica de *Accuracy* es calculada según la definición de la Ecuación (2):

$$Accuracy = \frac{\sum_{i=1}^n M_{ii}}{\sum_{i=1}^n \sum_{j=1}^n M_{ij}} \quad \text{Ecuación (2)}$$

donde, n es el número de dimensiones de la matriz M .

3. Visualizar: con el objetivo de analizar visualmente los datos y determinar que atributos pueden ser mejorados, en esta etapa el usuario analiza los datos de forma visual e interactiva. Para propiciar este tipo de análisis visual, en ésta etapa fue implementado el método de visualización de coordenadas paralelas (Gauthier-Villars, 1885), para lo cual los datos de todos los atributos fueron normalizados en el rango de [0-1]. Y con el fin de diferenciar las instancias de crímenes de diferente categoría, fueron asignados colores diferentes para cada categoría.
4. Reformular: después del análisis visual del conjunto de datos y cuando el usuario haya identificado los atributos a ser mejorados, el método vuelve a ser reiniciado en la etapa número (1). Cuando el método vuelve a ser reiniciado el usuario puede reformular los métodos de preprocesamiento y de creación de nuevos atributos. Esta etapa define un proceso iterativo, donde todo el método es ejecutado varias veces hasta que la calidad de clasificación deseada haya sido alcanzada.

Algoritmo 1. Algoritmo del método propuesto

```
Entrada: Conjunto de datos del crimen (DS).
1.   While(true): /*iteraciones del método*/
2.       Transformar ← DS;
3.       Preprocesamiento ← DS;
4.       Creación de nuevos atributos ← DS;
5.       Entrenar clasificador ← DS;
6.       Visualizar ← DS; /*interacción con el usuario*/
7.       If(clasificación es adecuada):
8.           Finalizar el algoritmo;
9.       End If
10.      Reformular ← DS;
11.      End While
Salida: Clasificador entrenado para la predicción del crimen.
```

Resultados

En la Figura (3) son ilustrados todos los resultados iniciales alcanzados con el método propuesto. La Tabla (2) muestra los resultados obtenidos con la agregación progresiva de los atributos *F6, *F7, *F8 y *F9. En los resultados se puede observar que para 30 categorías se alcanzó una calidad de clasificación del 91,00 %. Por otro lado, la Figura (4) ilustra la visualización del conjunto de datos generado para 10 categorías. La visualización del conjunto de datos de crimen es realizada a nivel de los atributos, por esta razón fue implementado el método de visualización de datos por coordenadas paralelas.

Figura 3. Resultados de clasificación: la figura ilustra 5 resultados de clasificación (02 - 30 categorías de crímenes) con atributos agregados progresivamente. Donde, F1-F5 son atributos procesados directamente desde el conjunto original, por otra parte *F6, *F7, *F8 y *F9 son 4 nuevos atributos creados.

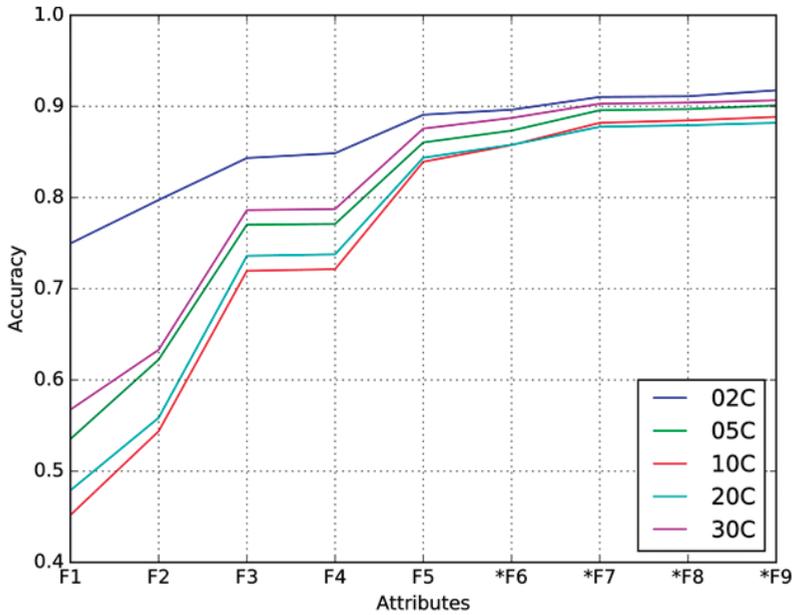
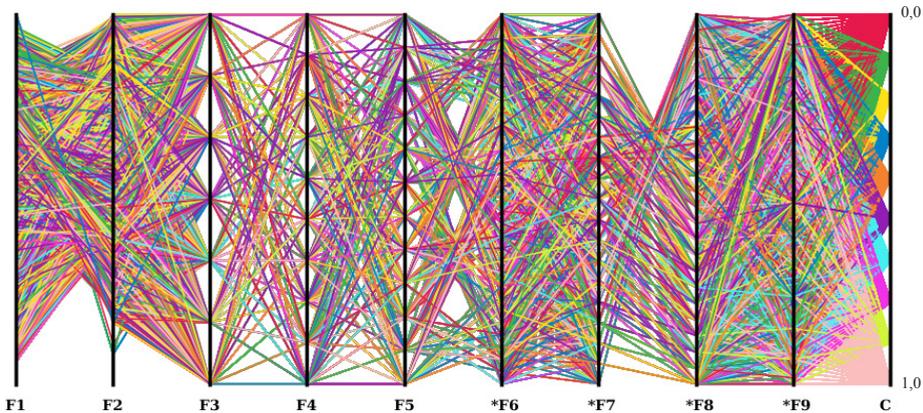


Tabla 2. Resultados de clasificación obtenidos con la agregación progresiva de los nuevos atributos creados *F6, *F7, *F8 y *F9.

Atributos	Categoría	Accuracy	Atributos	Categoría	Accuracy Ecuación (2)
*F6	02C	0,8963	*F8	02C	0,9111
	05C	0,8734		05C	0,8970
	10C	0,8575		10C	0,8846
	20C	0,8578		20C	0,8793
	30C	0,8873		30C	0,9041
*F7	02C	0,9103	*F9	02C	0,9176
	05C	0,8957		05C	0,9011
	10C	0,8820		10C	0,8885
	20C	0,8776		20C	0,8819
	30C	0,9029		30C	0,9067

Figura 4. Visualización del conjunto de datos generado: la figura ilustra los atributos visualizados empleando Coordenadas Paralelas. Donde, F1-F5 son atributos procesados directamente desde el conjunto de datos original, por otra parte *F6-*F9 son los 4 nuevos atributos creados, C es el atributo a predecir que corresponde a las categorías de crímenes, en este caso, son visualizadas 10 categorías de crímenes.



Discusión

Los resultados de la Figura (3) muestran un aumento progresivo de la calidad de clasificación, esto podría tener su origen por 2 razones. La primera razón, puede ser por los métodos empleados para el preprocesamiento y creación de los atributos, y la segunda razón podría ser por los procesos iterativos e interactivos del método propuesto. Se puede considerar iterativo porque el método es repetido varias veces, y se puede considerar interactivo porque el usuario interviene visualmente en la evaluación de la calidad de los atributos con el objetivo de reformular las etapas del método y de esta forma incrementar la calidad de clasificación.

Los 4 nuevos atributos *F6-*F9 están basados en datos tipo tiempo, porque se presume que existe una relación directa entre la actividad criminal y el tiempo, esto fue comprobado porque los nuevos atributos tipo tiempo incrementaron la calidad de clasificación en todos los resultados obtenidos. No obstante, aún no fueron explorados la creación de atributos basados en espacio (latitud y longitud), se podría suponer que al crear otros atributos basados en el espacio y al ser combinados con los atributos de tipo tiempo, la calidad de clasificación pueda ser incrementada aún más, esto en razón a que podría existir una relación directa entre tiempo, espacio y las categorías de crímenes.

En comparación a los métodos existentes que únicamente preprocesan los datos y entrenan clasificadores (Sathyadevan *et al.*, 2014; Vineeth *et al.*, 2016; Shamsuddin *et al.*, 2017), nuestro método se diferencia al incluir la participación interactiva del usuario en la clasificación de datos de crimen. Por otra parte, existen trabajos que emplean la visualización solo para visualizar instancias de los crímenes ya sea por mapas geográficos (Jayaweera *et al.*, 2015; Wang, 2015) o por mapas de calor (Wu *et al.*, 2017; Chae *et al.*, 2015), en cambio, el método propuesto emplea la visualización para mejorar la calidad de clasificación de datos del crimen durante el entrenamiento un clasificador.

Conclusiones

En este artículo fue formulado un método para predecir datos del crimen, este método está basado en el preprocesamiento de los datos, creación de nuevos atributos y la inclusión del usuario en el proceso para mejorar la calidad de clasificación durante el entrenamiento de un clasificador.

Una de las principales contribuciones en este trabajo es la inclusión del usuario donde participa de forma interactiva en los procesos iterativos del método propuesto. Los resultados iniciales del método propuesto alcanzaron una calidad de clasificación del 91,00 %. Futuros trabajos para mejorar los resultados de clasificación podrían incluir la creación de nuevos atributos basados en el espacio (latitud, longitud).

Referencias bibliográficas

- Brandoli, B., Eler, D., Paulovich, F., Minghim, R., & Batista, J. (2010). Visual data exploration to feature space definition. In 2010 23rd SIBGRAPI *Conference on Graphics, Patterns and Images*, 32–39.
- Chae, J., Wang, G., Ahlbrand, B., Gorantla, M. B., Zhang, J., Chen, S., Xu, H., Zhao, J., Hatton, W., Malik, A., Ko, S., & Ebert, D. S. (2015). Visual analytics of heterogeneous data for criminal event analysis vast challenge 2015: Grand challenge. In 2015 *IEEE Conference on Visual Analytics Science and Technology (VAST)*, 149–150.
- Nargesian, F., Samulowitz, H., Khurana, U., Khalil, E. B. & Turaga, D. (2017). Learning feature engineering for classification. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, 2529–2535.

- Gauthier-Villars (1885). *Coordonnées parallèles et axiales: Mthode de transformatio ngéométrique et procédé nouveau de calcul graphique déduits de la considération des coordonnées parallèles.*
- Gutierrez, D. (2015). *Machine Learning and Data Science: An Introduction to Statistical Learning Methods with R. Technics Publications, LLC.*
- Jayaweera, I., Sajeewa, C., Liyanage, S., Wijewardane, T., Perera, I., & Wi-jayasiri, A. (2015). Crime analytics: Analysis of crimes through news-paper articles. In *2015 Moratuwa Engineering Research Conference (MERCon)*, 277–282.
- Marsland, S. (2014). *Machine Learning: An Algorithmic Perspective*, (2nd ed.). Chapman & Hall/CRC.
- Sathyadevan, S., Devan, M. & Gangadharan, S. (2014). Crime analysis and prediction using data mining. In *2014 First International Conference on Networks Soft Computing (ICNSC2014)*, 406–412.
- Shamsuddin, N. H. M., Ali, N. A., & Alwee, R. (2017). An overview on crime prediction methods. In *2017 6th ICT International Student Project Conference (ICT-ISPC)*, 1–5.
- Vineeth, K. R. S., Pandey, A., & Pradhan, T. (2016). A novel approach for intelligent crime pattern discovery and prediction. In *2016 International Conference on Advanced Communication Control and Computing Technologies (ICACCCT)*, 531–538.
- Wang, D. (2015). Contrast pattern based methods for visualizing and predicting spatiotemporal events. In *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, 1560–1567.
- Wu, S., Male, J., & Dragut, E. (2017). Spatial-temporal campus crime pattern mining from historical alert messages. In *2017 International Conference on Computing, Networking and Communications (ICNC)*, 778–782.

E-mail: lizhh@usp.br

Recibido: 23 de setiembre de 2017

Aprobado: 2 de diciembre de 2017



Liz Maribel Huancapaza Hilasaca recibió el grado de bachiller en Ingeniería de Sistemas por la Universidad Andina Néstor Cáceres Velásquez, Perú, en 2009. Actualmente es estudiante especial en la Maestría de Ciencia de Computación y Matemática Computacional en el *Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo*, Brasil. Sus principales áreas de interés para investigación son ciencia de datos, visualización de información, aprendizaje máquina, predicción y visualización de datos espacio temporales.



Ivar Vargas Belizario recibió el grado de bachiller en Ingeniería de Sistemas por la Universidad Andina Néstor Cáceres Velásquez, Perú, en 2007. Recibió el grado de M.Sc. en Ciencia de Computación y Matemática Computacional por la *Universidade de São Paulo*, Brasil, en 2012. Actualmente es investigador y candidato al grado de doctor en Ciencia de Computación y Matemática Computacional en la *Universidade de São Paulo*, Brasil. Sus áreas de interés para investigación incluyen procesamiento de imágenes, reconocimiento de patrones, redes complejas, ciencias de datos, aprendizaje máquina y visualización de información.



Rodolfo Fredy Arpasi Chura doctor (c) en Ingeniería de Sistemas por la Universidad Nacional Federico Villareal. Profesor de Matemáticas, Universidad Nacional Mayor de San Marcos. M.Sc. en Informática y D.S. en Contabilidad y Administración por la Universidad Nacional del Altiplano, Ingeniero de Sistemas por la Universidad Andina Néstor Cáceres Velásquez. En la actualidad es profesor investigador en la Universidad Andina Néstor Cáceres Velásquez. Sus principales áreas de interés para investigación son ciencias de datos, neurociencias.