

JAIME ARBOLEDA CASTILLA

Científico de Datos | Matemático | Ingeniero Informático

Nacido en 1985 en Algeciras (Cádiz).

Me encanta obtener conocimientos a partir de los datos, y aprovecharlos para tomar decisiones mejores. Siempre estoy apasionado por aprender cosas nuevas. Estoy interesado especialmente en la utilización de la Inteligencia Artificial y el Machine Learning en el control del fraude tributario.

EXPERIENCIA PROFESIONAL

Actualidad
|
2024

Jefe de Área

Agencia Tributaria

Madrid

En la Subdirección de Tecnologías de Análisis de la Información en Investigación del Fraude del Departamento de Informática Tributaria

Descripción

- Funcionario del Cuerpo Superior de Sistemas y Tecnologías de la Administración del Estado.
- Jefe de Área de modelos predictivos en el control del fraude.

Proyectos

- Desarrollo de una plataforma ad-hoc para monitorizar los modelos de Machine Learning (MLOPS).
- Aplicación de redes neuronales convolucionales para detección de fraudes en imágenes de Rayos X en aduanas.

Tecnologías

- Python, Spark.
- Linux, Cloudera.
- Pandas, Numpy, Scikit-Learn, Xgboost, SyBase IQ.
- PyTorch, ZenML.



INFORMACIÓN DE CONTACTO

 github

 linkedin

 email

HABILIDADES Y CONOCIMIENTOS

- Sistema fiscal español.
- Herramientas de control del fraude tributario.
- Control aduanero.
- Machine Learning y Deep Learning.
- Análisis de datos y estadística.
- Python, R, Spark, Java.
- Numpy, Pandas, Scikit-Learn, PyTorch, Tensorflow, Keras.

2023
|
2021

Data Scientist

Comisión Europea

Trabajo remoto

En el proyecto Safety and Security Analytics (SSA) ICS2 Project, TAXUD

Descripción

- Investigación, diseño y desarrollo continuo de modelos en SSA.
- Apoyo en la implementación y orquestación de modelos en SSA, y soluciones analíticas para uso en tiempo real.
- Apoyo en el diseño, organización y supervisión de todo el flujo de trabajo en SSA.
- Desarrollo de capacidades de búsqueda y detección de anomalías.
- Formación a funcionarios de todos los Estados Miembros.

Proyectos

- Motor de búsqueda de similitud (entre declaraciones de aduanas) usando redes neuronales.
- Herramienta de comparación de modelos para análisis de rendimiento, construida con Dash y Docker.
- Generación sintética de XMLs.
- Herramienta de procesamiento de XMLs.
- Detección de Anomalías en las declaraciones de aduanas.

Tecnologías

- Dataiku, RStudio, JupyterLab.
- Denodo, Oracle, PostgreSQL, Neo4j.
- SQL, Python, R, Bash, XML, JSON, HTML.
- Numpy, Pandas, Keras, Tensorflow, Dash.
- GitLab, Jenkins, Docker, Kubernetes, Apache Kafka.

2021
|
2020

Jefe de Área

Agencia Tributaria

Madrid

En la Subdirección de Tecnologías de Análisis de la Información en Investigación del Fraude del Departamento de Informática Tributaria

Descripción

- Funcionario del Cuerpo Superior de Sistemas y Tecnologías de la Administración del Estado.
- Jefe de Área de modelos predictivos en el control del fraude.

Proyectos

- Algoritmo de clustering personalizado (basado en KNN) que, utilizando la información de compras y ventas de cada empresa, predice si su sector declarado de actividad económica es correcto o no.
- Identificación de la posición estadística y subvaloración de bienes en declaraciones aduaneras mediante la modificación de un algoritmo existente proporcionado por la Comisión Europea.
- Clasificador (utilizando XGBoost) para predecir cuándo es más probable que un contribuyente cometa un error al modificar algunas partes de su borrador de IRPF. El objetivo era enviar un mensaje de advertencia a estos contribuyentes en caso de modificación, con el fin de reducir errores.
- Clasificador para predecir el riesgo de impago de deudas con la Agencia Tributaria, con el propósito de anticipar medidas preventivas.
- Clasificador para predecir el riesgo de no pagar sus obligaciones fiscales a tiempo para un contribuyente dado. Este modelo utiliza información casi en tiempo real sobre todas las facturas recopiladas en los meses anteriores a la predicción.
- Modelo de regresión para predecir los ingresos totales (declarados o no declarados) de una familia dada utilizando toda la información disponible.

Tecnologías

- Python, Scala, Spark.
- Linux, Cloudera.
- Pandas, Numpy, Scikit-Learn, Xgboost, Luigi.
- SyBase IQ, DataStage.

2019
|
2017

Jefe de Área

Agencia Tributaria

Madrid

En la Subdirección de Aplicaciones del Departamento de Informática Tributaria

Descripción

- Funcionario del Cuerpo Superior de Sistemas y Tecnologías de la Administración del Estado.
- Jefe de Área de Aplicación Gestora de IRPF.

Proyectos

- Servicio web para la ingesta de datos personales, para la aplicación Renta Web.
- Servicio criptográfico para otorgar credenciales de acceso para la presentación de las Declaraciones de IRPF.
- Ingesta de datos de Declaraciones de IRPF.
- Análisis de riesgos (combinando riesgos basados en reglas, riesgos estadísticos y modelos predictivos simples) para las Declaraciones de IRPF.
- Desarrollo de software para la gestión y ciclo de vida de las Declaraciones de IRPF.

Tecnologías

- COBOL, Java, HTML, JavaScript.
- Web Services.
- DB2, Oracle.
- Z/OS, Linux.
- SyBase IQ, DataStage.

2017
|
2013

Jefe de Servicio

Agencia Tributaria

Madrid

En la Subdirección de Aplicaciones del Departamento de Informática Tributaria

Descripción

- Funcionario del Cuerpo Superior de Sistemas y Tecnologías de la Administración del Estado.
- Jefe de Área de Aplicación Gestora de Sociedades

Projects

- Ingesta de datos de Declaraciones de Sociedades
- Análisis de riesgos (combinando riesgos basados en reglas, riesgos estadísticos y modelos predictivos simples) para las Declaraciones de Sociedades.
- Desarrollo de software para la gestión y ciclo de vida de las Declaraciones de Sociedades.

Technologies

- COBOL, Java.
- DB2, Oracle.
- Z/OS, Linux.
- SyBase IQ, DataStage.



EXPERIENCIA DOCENTE

2023

Big Data

BBVA

📍 Trabajo Remoto

Profesor de Data Scientist Fundamentals, impartido a 23 trabajadores del BBVA en Mexico, Argentina y Colombia. El curso tuvo una duración de 42 días (168 horas), y cubrió entre otros:

- Big Data tools en BBVA (Datio, Stratio, Crossdata).
- Python.
- Data Wrangling con Numpy y Pandas.
- Data Visualization con Matplotlib y Seaborn.
- Machine Learning con scikit-learn.
- Big Data con Spark y SparkSQL.
- Machine Learning con SparkML.
- Deep Learning con PyTorch.

2021

Seminario

Webinar

📍 Universidad Complutense, Madrid

Di una charla en el webinar sobre Fiscalidad e Inteligencia Artificial, con Ramón Palacios (Subdirector del Departamento de Verificación y Control Tributario) sobre el proyecto Nudge para impulsar el control tributario adelantándolo a la fase de Asistencia al Contribuyente.



PUBLICACIONES

2021

Proyecto Nudge

Paper

📍 Aranzadi Thomson Reuters

El proyecto Nudge, realizado en la Agencia Tributaria, consistió en la aplicación de la Inteligencia Artificial para ayudar en la asistencia al contribuyente y el cumplimiento voluntario de las obligaciones fiscales. Fue publicado en Aranzadi Thomson Reuters, junto con otros trabajos presentados en el seminario web "Fiscalidad e Inteligencia Artificial" organizado por la Universidad Complutense de Madrid.



COLABORACIONES EN PROYECTOS OPEN SOURCE

2023

Colaborador de category_encoders

category_encoders

📍 Trabajo Remoto

Arreglé un error relacionado con la compatibilidad entre la librería y `sklearn`. Mi solución fue integrada en el proyecto.

2022

Colaborador de Keras

Keras

📍 Trabajo Remoto

Encontré un error, y semanas después pude resolverlo con una solución que fue integrada en el proyecto.

2021

Desarrollador de una librería Open Source

Nested Cross Validation

📍 Trabajo remoto

Librería de Python que hace hyperparameter optimization y probability calibration sobre modelos de clasificación usando un enfoque de Nested Cross-Validation.



FORMACIÓN


2023	DataTalksClub Zoomcamp	Remoto
• Data Engineering Zoomcamp		
2022 2021	Comisión Europea Formación interna	Remote
• Cibersecurity		
• Software Development y Agile Methodologies.		
2021 2013	Agencia Tributaria Formación interna	Madrid
• Análisis de Datos.		
• Machine Learning y Big Data.		
• Geospatial Data Processing en R.		
• Metodologías Ágiles.		
• Zújar (herramienta interna de BI).		
• Genio (herramienta interna de reporting).		
• Blockchain.		
• OSGI y Java.		
2022 2017	Coursera Cursos y especializaciones	Remote
• Probabilistic Graphical Models: Representation		
• Probabilistic Graphical Models: Inference		
• Probabilistic Graphical Models: Learning		
• Bayesian Statistics: From Concept to Data Analysis.		
• Bayesian Statistics: Techniques and Models.		
• Bayesian Statistics: Mixture Models.		
• Neural Networks and Deep Learning.		
• Improving Deep Neural Networks: Hyperparameter Tuning, Regularization and Optimization.		
• Structuring Machine Learning Projects.		
• Convolutional Neural Networks and Computer Vision.		
• Sequence Models and Natural Language Processing.		
• Machine Learning.		
2013 2012	INAP Curso selectivo	Madrid
Curso de acceso al Cuerpo Superior de Sistemas y Tecnologías de la Información del Estado.		
• Aprobado con la segunda mejor calificación de todos los candidatos.		
2016 2014	Universidad Española de Educación a Distancia (UNED) Master en formación del profesorado	Madrid
• Nota media de 8.1.		
• Finalizado con un trabajo de enseñanza de matemáticas usando programación.		
2010 2009	Universidad Complutense de Madrid (UCM) Master en Investigación Matemática	Madrid
• Nota media de 8.8.		
• Finalizado con un trabajo de investigación en cuantización geométrica.		

2009
|
2005



Universidad Autónoma de Madrid (UAM)

Double grado en Matemáticas e Informática

 Madrid

- Nota media de 9.5.
- 25 Matrículas de Honor.
- Premio al mejor estudiante de la promoción.