

Mabel Villalba Jiménez

Data Engineer & Team Lead

Scalable Data Solutions, Machine Learning, and Agile Leadership

mabelvj@gmail.com

LinkedIn GitHub Website StackOverflow

skills	experience
<p>Cloud Computing: AWS (Lambda, Glue, S3, CloudFormation, Kinesis), Microsoft Azure</p> <p>Agile & Project Management: Agile, Scrum, Team Leadership, Stakeholder Engagement, Roadmap planning</p> <p>Leadership: Team Management, Mentoring, Career Development, Performance Review</p> <p>Programming: Python, R, SQL, Unit Testing</p> <p>Data Engineering: ETL Pipelines, Pandas, PySpark, MongoDB</p> <p>AI & Machine Learning: Scikit-learn, TensorFlow, Deep Learning, Time Series Analysis</p> <p>Software Development: Docker, Docker Compose, CI/CD</p>	<p>Data Engineer SEAT:CODE (April 2021 - Present)</p> <p>Designed and maintained cloud-based data infrastructures, supporting large-scale analytics.</p> <p>Developed and deployed real-time ETL workflows using AWS Lambda, Kinesis, Glue, and S3.</p> <p>Optimized SQL performance and data modeling, fine-tuning queries for efficiency.</p> <p>Established data validation processes and governance frameworks. Integrated CI/CD pipelines to automate deployment and testing of data workflows.",</p> <p>Enhanced system monitoring and reliability by implementing logging and alerting solutions.</p> <p>Implemented predictive maintenance solutions using machine learning models and developed AWS Quicksight dashboards por visualization, providing real-time insights.</p> <p>Team Lead SEAT:CODE (May 2023 - November 2024)</p> <p>Led a multidisciplinary team of data engineers, data scientists, facilitating cross-team discussions and agreements.</p> <p>Owned the full lifecycle of PIBOD, collaborated with the Product Owner to defining its roadmap and technical feasibility and coordinated tis ramp-down.</p> <p>Conducted 1:1 mentoring sessions, career planning, and guided team members using Profile Ladders, as well as executed performance evaluations</p> <p>Coordinated the definition and tracking of team OKRs, along with the Business Unit's.</p> <p>Participated in hiring processes, ensuring a diverse and talented team.</p> <p>Data Science Engineer Mática Partners (December 2020 - April 2021)</p> <p>Designed and managed scalable ETL pipelines using PySpark and Docker, processing datasets with millions of records efficiently.</p> <p>Deployed solutions on AWS and Azure, achieving seamless cloud integration for clients.</p> <p>Automated CI/CD workflows with Jenkins, reducing deployment times by 40%.</p> <p>Python Developer Zyte (April 2019 - December 2020)</p> <p>Developed advanced web scraping solutions using Scrapy, enabling high-volume data extraction for Fortune 500 clients.</p> <p>Optimized ETL pipelines with Pandas and SQL, improving data processing speeds.</p> <p>Integrated Celery for distributed task management, processing over 1M tasks daily.</p> <p>Implemented robust monitoring and debugging systems with Sentry.</p> <p>Data Scientist and Python Developer Independent Contractor (May 2017 - April 2019)</p> <p>Developed algorithmic trading systems on Quantopian for a private fund.</p> <p>Created a TV attribution and ROI reporting system for Windsor AI, leveraging R and PostgreSQL for real-time insights.</p> <p>Built scalable web scrapers for LISTedTECH, automating updates for large educational databases.</p> <p>Mentor and Reviewer Udacity (April 2017 - May 2019)</p> <p>Mentored over 120 students in Data Science and Machine Learning.</p> <p>Reviewed 750+ projects across Deep Learning and Reinforcement Learning, maintaining a 4.93/5 feedback rating.</p> <p>Predoctoral Researcher UPC (October 2015 - October 2016)</p> <p>Designed Python and MATLAB simulations for optical devices, achieving a 54dB rejection on sidebands for optical networks.</p> <p>Contributed to the development of wavelength shifters for next-generation optical communication systems.</p> <p>education</p> <p>MSc in Photonics Polytechnic University of Catalonia (UPC)</p> <p>Telecommunication Engineering (BSc + MSc) University of Malaga</p> <p>Data Analyst Nanodegree Udacity</p> <p>Machine Learning Engineer Nanodegree Udacity</p> <p>Machine Learning Course Stanford University, Coursera</p>

Projects and Contributions

Python

pandas

BUG: Groupby quantiles incorrect bins #33200 - Fixed an issue where incorrect bins were being calculated for quantiles when using groupby, improving statistical accuracy.

BUG: Add unordered option to pandas.cut (#33141) - Added support for unordered categories in pandas.cut, enhancing its flexibility for categorical data.

BUG: pivot_table not returning correct type when margin=True and aggfunc='mean' - Resolved inconsistencies in pivot_table results with margins enabled, ensuring type correctness.

DOC: update pandas.DataFrame.boxplot docstring. Fixes #8847 - Improved documentation for pandas.DataFrame.boxplot, providing clearer usage examples.

scikit-learn

[MRG+1] Fixes #10393 Fixed error when fitting RidgeCV with integers - Resolved an error in RidgeCV fitting with integer inputs, increasing model reliability.

[MRG+1] Fixes #10284 Added store_cv_values to RidgeClassifierCV and a test - Introduced a new feature to store cross-validation values in RidgeClassifierCV, aiding model diagnostics.

Fixes #10216 export_graphviz should work with a sklearn.tree._tree.Tree - Enhanced compatibility for exporting decision trees with export_graphviz.

scrapy

[MRG+1] Itemloader errors - Fixed errors in item loaders, ensuring consistent data extraction.

Raise error when start_url found instead of start_urls #4170 - Added error handling for incorrect start_url usage, improving user feedback.

shub

[hide-apikey] Fixes #264: Hide apikey - Implemented a feature to hide API keys, enhancing security during operations.

Insecure exception with credential should raise warning #355 - Added warnings for insecure exceptions involving credentials, promoting safer practices.

StackOverflow Highlights

28 votes Groupby sum and count on multiple columns in Python

Feb 13, 2018

Resolved a common issue of applying groupby operations with both sum and count across multiple columns in pandas. The solution was praised for its clarity and efficiency, helping many developers streamline their data analysis workflows. Comments highlighted how well-structured and practical the explanation was.

12 votes Adding value to only a single column in pandas DataFrame

Nov 15, 2018

Provided an elegant approach to modifying a single column in a pandas DataFrame without affecting others. This solution stood out for its simplicity and efficiency, addressing a frequent need in data manipulation tasks. Feedback highlighted the helpfulness of the provided code snippets.

12 votes Cannot make seaborn violin plot horizontal [Python3.X]

Feb 14, 2018

Offered a detailed solution to plot horizontal violin plots using seaborn, clarifying an often-misunderstood aspect of the library. This answer included well-commented examples, making it easier for users to adapt the solution to their needs. Comments appreciated the step-by-step approach.

11 votes Print the decision path of a specific sample in a random forest classifier

Feb 20, 2018

Demonstrated how to extract and print the decision path of a sample in a random forest classifier, addressing a nuanced use case in machine learning. The solution was praised for bridging theoretical concepts with practical implementation. Comments lauded the clarity and precision.

8 votes How to resample 6 months

Sep 27, 2018

Provided a concise and effective solution for resampling data into 6-month intervals using pandas. This answer addressed a niche but recurring problem in time-series data analysis. Users noted how the explanation clarified an otherwise complex operation.