David Vargas Leos

+52 (449)257-9204 | davale1201@gmail.com |

SUMMARY

Data Scientist and Software Engineer with 2 years of real world consultancy experience. Experienced in framing business problems into quantitative and computational tasks, and translating these requirements into efficient, production ready code for data analysis, experimentation, and deployment. Spanish Native, English (C1), French (B2).

SKILLS

Coding Languages: Python, C++, SQL, Bash.
Infrastructure & Tools: Docker, Linux, SLURM, Git, MLFlow, Kedro, Snowflake, AWS, PySpark, PostgreSQL.
Libraries: PyTorch, Pandas, NumPy, SciKit-Learn, Polars.
Systems & Networking: Linux internals, TCP/IP fundamentals, client server architecture, RESTful services.
Certifications: SnowPro Associate: Platform, AWS Solutions Architect Associate (in progress).

EDUCATION

Instituto Tecnológico de Monterrey (#1 in Mexico, QS Rankings)
B.S. in Data Science and Mathematics (GPA: 93/100)   *Graduated: July 2025.*
- **Courses:** Data Structures & Algorithms, Deep Learning, MLOps, Probability, Multivariate Statistics, Discrete & Stochastic Optimization, Topology Data Analysis, Cryptography, Control Theory.
- Scholarship for Academic Excellence. Member of Academic Group IAnswers.

EXPERIENCE

Data Science Professor | Aurora University (Remote)   September 2025 - Present
- **Probability:** Taught probability distributions, combinatorial methods applied to data science problems.
- **Discrete Mathematics:** Taught core CS foundations including algorithm complexity, graph and tree algorithms.
- **Data visualization:** Taught image processing and data visualization techniques, covering 2D/3D transformations and statistical graphics libraries (Matplotlib, Seaborn, ggplot2) for exploratory analysis and model visualization.

Computational Researcher | Purdue University (#88 Globally, QS Rankings)   August 2024 - December 2024
*Supervisor: Dr. Michael Gribskov (**H-index:52**)*
- Researched RNA secondary structure prediction, building computational tools to support biological function discovery.
- Engineered high throughput High Performance Computer (**HPC**) pipeline using **SLURM** job scheduling and graph isomorphism algorithms, processing **10,000+ RNA structures**.

CONSULTANCY EXPERIENCE

MLOps Consultant | *TCA Software*   February - June 2025
- Engineered a modular time series forecasting pipeline using **Kedro** and **XGBoost** to predict occupation levels, ensuring full reproducibility via **MLflow** experiment tracking and **Docker** containerization.
- Deployed an interactive **Streamlit in Snowflake (SiS)** dashboard, enabling stakeholders to visualize model outputs in real time and optimize decision making.

Data Scientist Consultant | *GeoStats*   August 2023 - June 2024
- Built a geospatial crime prediction system using **Random Forest** and **GeoPandas**, integrating demographic census data across **50+ neighborhoods** to identify 12 high-vulnerability zones.
- Collaborated with the **Government of San Pedro Garza García** to optimize pollution detector placement by solving the **Maximal Covering Location Problem (MCLP)** using **PuLP**, presenting strategic insights via **Looker Studio**.

Environmental Data Consultant | *Secretaría de Medio Ambiente*   Augusts - December 2023
- Engineered a pollution classification model for the **Secretaría de Medio Ambiente** using data from **real world environmental sensors**, applying **PCA** and **Logistic Regression** to achieve **82% accuracy** in PM2.5 detection.

Operations Research Consultant | *Government of Nuevo León*   June 2023
- Formulated a multi objective route optimization model for the **Government of Nuevo León** to generate optimal itineraries, solving the **Orienteering Problem with Time Windows (OPTW)** subject to budget and user rating constraints.
- Implemented the optimization solver using **GAMS** and integrated the **Bing Maps API**, validating the theoretical routes against real world geospatial distances and travel times.

Data Science Consultant | *Ternium*   June 2023
- Engineered an **NLP** and **K-Means clustering** pipeline to automate the classification of industrial failure reports, successfully handling heterogeneous text and user generated spelling errors.
- Standardized the maintenance database structure, significantly reducing manual analysis time and enabling faster root cause identification for engineering teams.