# Claudio **Scalzo**

Software Engineering and Data Science student • Data Engineer intern at SAP

(+39) 346 24 55 116 • claudio.scalzo@outlook.com • github.com/claudioscalzo • linkedin.com/in/claudioscalzo

*Born in Italy in 1994. Passionate about software development in the big data landscape. Strong analytical thinking and team-working skills. Successful results due to the passion in this field and the high precision in projects and tasks fulfillment. Languages: Italian (MT), English (C1), French (intermediate).*

## Education

### Master of Science in **Data Science and Engineering**

*Sophia Antipolis, France*

Eurecom / Telecom ParisTech

*Sep. 2017 - Mar. 2019*

- Double Degree program between Telecom ParisTech and Politecnico di Torino

### Master of Science in **Software Engineering**

*Turin, Italy*

Politecnico di Torino

*Sep. 2016 - Mar. 2019*

### Bachelor's Degree in **Software Engineering**

*Turin, Italy*

Politecnico di Torino

*Sep. 2013 - Jul. 2016*

- with the highest grade, 110/110

## Experience

### SAP France

*Paris, France*

Data Engineer intern

*Jul. 1, 2018 - Dec. 31, 2018*

- Worked on the SAP Mass Data Extension team, adding big data transfer capabilities, external data ingestion and master data matching in the company Azure Data Lake, built to extend data warehousing functionalities offered by SAP HANA.
- Worked on Python code leveraging Spark parallelization and ML techniques, exploiting the distributed computing power offered by the HDInsight clusters and ADLS storage. Worked under the Agile Software Development methodology.
- Managed Jenkins pipelines for both testing and production environments, providing help and support during the daily team operations, troubleshooting and solution proposals.

## Projects

### Team leader for an optimization project for TIM / SWARM Joint Open Lab

GITHUB.COM/CLAUDIOSCALZO/COIOTE

- Solving of a *VRP (Vehicle Routing Problem)* optimization problem proposed by *TIM* and the *SWARM Joint Open Lab*, for an IOT project named *CoIoTe*. Multistart tabu-search approach, written in Java (with the *OpenTS* Java library).
- Achieved 1st position in the final ranking. Secured great comprehension of metaheuristics. Improved algorithmic and team-working skills.

### Virtual Assistant for answering music related questions

GITHUB.COM/D2KLAB/MUSIC-CHATBOT • CHATBOT.DOREMUS.ORG

- Developing of a virtual assistant (in the chatbot and vocal assistant forms), capable of answering music related questions and providing detailed graphical results. Informations extracted from the *DOREMUS* knowledge base, queried using the *SPARQL* language.
- Built with *Node.js*, using the *BotKit* framework. Trained Google's *Dialogflow* as NLP. *Facebook Messenger*, *Slack* and *Google Assistant* support.
- Contributed with two accepted pull requests to the `botkit-middleware-dialogflow` author, for concurrency and language support.

### House prices *Kaggle* challenge: predicting sales prices with advanced regression techniques

GITHUB.COM/LOMLUCA/AML

- Solution of the known Kaggle challenge. Achieved top grade on the course ranking thanks to smart preprocessing techniques (like PCA and DBSCAN for outlier removal), and stacked tree-based and regularized regression models.
- Written in *Python*, using *Pandas DataFrames* for the data structures and *scikit-learn* for the modeling phase.

### Challenge on the `Fashion-MNIST` and `CIFAR-10` datasets: Naive Bayes Classifier and Bayesian Linear Regression

GITHUB.COM/CLAUDIOSCALZO/ASI-CHALLENGE

- Solution of the *ASI (Advanced Statistical Inference)* course challenge. Implemented (from scratch) the Naive Bayes Classifier and the Bayesian Linear Regression, exploited in the classification tasks of the Fashion-MNIST and CIFAR-10 images datasets.
- Written in *Python* using *NumPy* and *Pandas*. Achieved extremely satisfying results in terms of accuracy and computational efficiency.

## Skills

| | |
|---|---|
| **Languages** | Python (+ PySpark, NumPy, Pandas, scikit-learn, Keras) • C • Java • Oracle PL/SQL |
| **Big Data** | Apache Spark • MapReduce & HDFS • NoSQL Architectures • Data Processing (cleansing, analysis) • Distributed Cluster Computing (Microsoft Azure: HDInsight, ADLS) • Jenkins CI/CD • NLP Techniques |
| **Machine Learning** | Deep Learning (NNs, CNNs, RNNs) • Probabilistic Machine Learning (Bayesian Classification, Regression, Mixture Models) |
| **OS & Other Tools** | GNU/Linux (+ Bash, AWK, Sed) • Git • Jupyter / Zeppelin Notebooks • Dialogflow (+ BotKit) • MATLAB |