Create knowledge.
Foster change.

UTAustin Portugal

Advanced Computing Training Program

Final Report

Participant Name:	Pedro Gabriel Dias Ferreira
Affiliation and Position:	Assistant Researcher at Instituto de Patologia e Imunologia Molecular da Universidade do Porto and Instituto de Investigação e Inovação da Universidade
Period:	20/09/2018 – 5/12/2018

During my stay at TACC/UTIAustin I integrated the Data Science group within the Women's Health Data Research Program. The group is a joint initiative of researchers from three institutes: The Institute of Computational Engineering and Sciences (ICES), The Texas Advanced Computing Center (TACC) home to the largest open science supercomputer in the world, and The Department of Statistics and Data Science. The group takes advantage of the world-class expertise in computational sciences at the University of Texas at Austin campus and the interaction with the medical doctors at Dell Medical School.

The main mission of the group is to develop computational tools for the prediction disease risk or adverse health outcome at an individual basis. By following the line of personalized medicine it aims to improve health outcomes and reduce the risks of healthcare. The group has a focus on pregnancy since it is a process that occurs in a relatively short time frame with frequent monitoring for which adequate medical actions can be taken. Therefore, pregnancy provides a good opportunity to develop individualized models for health care.

Within the line of the group my first task was to understand and explore a large and multidimensional dataset with hundreds of features and millions of records, across several years, of pregnancies statistics obtained throughout the United States. In the second phase, I have developed Machine Learning methods for classification and regression of different aspects of the pregnancy process.

Within this process I developed novel skills on modeling prediction from a high-dimensional dataset. In particular, skills in the optimization of the model to run millions of instances were developed. Moreover, I learn how to use TACC User Portal to run R and Python applications in TACC clusters.

Within this process we learned the difficulties and challenges of using datasets with dozens of millions of entries and dozens of features to predict several variables of the pregnancy process. Different optimization techniques were applied and tested in order to improve the performance of machine learning models. We have achieved prototypes of Machine Learning models with interesting results in their performance for prediction of pregnancy variables.

With the experience acquired during this training, when back in Portugal, I have been applying different Machine Learning methodologies for the prediction of clinical phenotypes from gene expression data, also from individualized datasets. I hope to keep the contact and collaboration with the group at UTAustin and apply jointly for future project calls.

This training experience at TACC/UTAustin allowed me to develop new computational skills and machine learning methodologies to be able to analyze and create prediction models for very large and complex datasets. This experience is now being applied actively in many aspects of my current work.