Capstone Project ideas for Data Science Track

Project 1: To predict survival outcome of Titanic Disaster

Goal: To predict survival outcomes from the 1912 Titanic disaster based on each passenger's features, such as sex and age. One can start with a simple algorithm and increase its complexity until model is able to accurately predict the outcomes for at least 80% of the passengers in the provided dataset.

https://www.kaggle.com/c/titanic

Project 2: Somatic Germline Mutation: Personalized Medicine

Rationale: The number of somatic germline mutations in human genes associated with inherited and acquired diseases is increasing exponentially. Yet, we do not have a comprehensive collection of these in the form of a freely available resource for easy mining and discoveries. This proposal is an effort to create such an invaluable resource that will be made available to all scientists.

Goal: To create a one-stop-shop comprehensive collection of mutation data for easy discovery (classification/prediction) in the era of personalized medicine.

Data: The underlying data will be collected from the following resources and literature mining.

1. Uniprot.org
2. OMIM
3. Clinvar
4. dbsnp
5. Cosmic
6. Cbio portal
7. GWAS
8. Text mining (pubmed)

Project 3: Heterogeneity Activity Recognition Data Set

Goal: Is to correctly classify activity in 6 categories based on data. The Heterogeneity Human Activity Recognition (HHAR) dataset from Smartphones and Smartwatches is a dataset devised to benchmark human activity recognition algorithms (classification, automatic data segmentation, sensor fusion, feature extraction, etc.) in real-world contexts; specifically, the dataset is gathered with a variety of different device models and use-scenarios, in order to reflect sensing heterogeneities to be expected in real deployments

http://archive.ics.uci.edu/ml/datasets/Heterogeneity%20Activity%20Recognition

Project 4: Heart Disease Data Set

Goal: Is to correctly classify if a patient has Heart Disease based on various attributes. The "goal" field refers to the presence of heart disease in the patient. It is integer valued from 0 (no presence) to 4. Experiments with the Cleveland database have concentrated on simply attempting to distinguish presence (values 1,2,3,4) from absence (value 0).

http://archive.ics.uci.edu/ml/datasets/Heart+Disease

Project 5: Credit Card Fraud Detection

Goal: To classify if the given credit card transactions is fraudulent or genuine. The datasets contain transactions made by credit cards in September 2013 by European cardholders. This dataset presents transactions that occurred in two days, it contains 492 frauds out of 284,807 transactions.

https://www.kaggle.com/dalpozz/creditcardfraud/data