

# PM 272 – Introduction to Health Data Science

## *End-of-term Assignment*

This assignment is worth **70% of your grade for this module** and is designed to test your R, Bioinformatics and Machine Learning knowledge.

This assignment is released on 28th November 2022 and is due, via Turnitin, on 3<sup>rd</sup> January 2023.

### Question

There are 3 data sources for this assignment:

- Humvar – a list of variants with known impact (Benign/Pathogenic) across all diseases
- Lab Variants – a list of mutations in HGVS format and the phenotype the patients had
- Variants Annotated – The mutations in Lab Variants, annotated with Variant Effects.

You must train a **Decision Tree** using the Humvar dataset, and report the **Accuracy, True Positive Rate and False Positive Rate** using a Train/Test experiment. Then, using the trained humvar decision tree, **predict the outcome of mutations in epilepsy and muscular conditions**, again reporting Accuracy, True Positive Rate and False Positive Rate.

Is it easier to predict when a mutation will cause epilepsy compared to predicting if a mutation will cause muscular conditions? How can we tell?

Finally, plot Receiver Operating Curves (ROC) for each disease and explain the interpretation of the ROCs.