**MSc Project Report**

16th September 2019

John H Booth (13133420)

BBK\_BUCI058D7\_1819 - MSc Data Science Project

MSc in Data Science

Department of Computer Science and Information Systems, Birkbeck College, University of London, 2019

Identifying Feature Importance in Pediatric Post-Mortem Outcome with Machine Learning Models

Academic Declaration

I have read and understood the sections of plagiarism in the College Policy on assessment offences and confirm that the work is my own, with the work of others clearly acknowledged. I give my permission to submit my report to the plagiarism testing database that the College is using and test it using plagiarism detection software, search engines or meta-searching software.

Contents

[1 Abstract 4](#_Toc15576391)

[2 Introduction 5](#_Toc15576392)

[3 Data 6](#_Toc15576393)

[4 Methods 7](#_Toc15576394)

[5 Analysis 8](#_Toc15576395)

[6 Results 9](#_Toc15576396)

[7 Conclusion 10](#_Toc15576397)

# Abstract

Post mortems are complex procedures that utilise a significant amount of hospital resources, yet despite this, cause of death is only determined in 45% of cases. The event itself can be very traumatic for the parents of the child, yet is essential for providing further clinical understanding of the patient’s cause of death. Given this, there is an imperative to extract the greatest possible value from the data. Here, we investigated whether machine learning could be used to derive novel insights from the prediction of post mortem outcomes.

A post mortem database containing 7000 records across 300 variables was analysed and categorised into stage of examination (external and internal). The outcome of the examination was summarised as either ‘cause of death determined’ or ‘not determined’. From these summarised data, cases were filtered by children aged <= 2 years, resulting in a dataset of 3,100 post mortems.

Following this, decision tree, random forest, and gradient boosting machine models were iteratively built for each stage of the post-mortem examination and compared using their accuracy metrics.

The naïve decision tree model using external examination data had a predictive performance of 67%. Model performance notably increased when trained on internal examination data. At each stage of the examination, a core set of data items, of which the final set included age, BMI, and heart weight were highlighted using model feature importance as key variables for determining post mortem outcome. The use of increasingly complex modelling techniques was able to boost the predictive performance of the model by as much as 10%.

This project clearly shows the value of collecting clinical procedural data which can then be modelled using machine learning techniques to inform clinical practice. With more time, further modelling, including unsupervised clustering could be undertaken to derive further insights.

# Introduction

## Background

## Literary Review

# Data

## Source

## Data Cleaning

* The nature of the data is that it is not normal

## Missing data

* Gestation in days at birth.

## Normalisation vs Standardisation

* Z score
* Normalisation for Age
* One-hot encoding

# Methods

## Create models

## Tune models

# Analysis

# Results

# Conclusion

# References

Archer, K.J., 2010. rpartOrdinal: an R package for deriving a classification tree for predicting an ordinal response. *Journal of Statistical Software*, *34*, p.7.

Luellen, J.K., Shadish, W.R. and Clark, M.H., 2005. Propensity scores: An introduction and experimental test. *Evaluation Review*, *29*(6), pp.530-558.

(Example of using RPART package)

Therneau, T.M. and Atkinson, E.J., 1997. An introduction to recursive partitioning using the RPART routines.

# Appendix A

# Random Thoughts

* Python vs R
  + Why used when
* Reproducible results
* R code; evolved through three stages
  + Individual model
  + Model run for all PM stages
  + Model as a function called with multiple random keys.