**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.

The report may be freely copied and distributed provided the source is explicitly acknowledged.

Acknowledgements

Contents

[1 Abstract 4](#_Toc16057319)

[2 Introduction 5](#_Toc16057320)

[2.1 Background 5](#_Toc16057321)

[2.2 Literary Review 5](#_Toc16057322)

[3 Data 6](#_Toc16057323)

[3.1 Source 6](#_Toc16057324)

[3.2 Data Cleaning 6](#_Toc16057325)

[3.3 Missing data 6](#_Toc16057326)

[3.4 Normalisation vs Standardisation 6](#_Toc16057327)

[4 Methods 7](#_Toc16057328)

[4.1 Create models 7](#_Toc16057329)

[4.2 Tune models 7](#_Toc16057330)

[5 Analysis 8](#_Toc16057331)

[6 Results 9](#_Toc16057332)

[7 Conclusion 10](#_Toc16057333)

[References 11](#_Toc16057334)

[Appendix A 12](#_Toc16057335)

[Random Thoughts 13](#_Toc16057336)

# 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.

Supervisor:

Nigel Martin (nigel@dcs.bbk.ac.uk)

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

# Introduction

## Background

Great Ormond Street Hospital for Children NHS Trust (GOSH) is the country’s leading centre for treating sick children. With the UCL Great Ormond Street Institute of Child Health, GOSH is the largest centre for paediatric research outside the US.

Specialist Paediatric Pathologists perform perinatal, infant and childhood post-mortems including hospital referrals, forensic cases and those on behalf of Her Majesty’s Coroner.

The Pathology Department has established a research database containing details of all post-mortems performed between 1996 and 2017.  The database was originally used specifically for research into Sudden Unexpected Death in Infancy (SUDI). Since then it has been utilised for a number of other projects investigating SUDI, stillbirths and various aspects of paediatric autopsy procedure.

Currently the database holds 7000 records, each record representing an individual post-mortem. Up to 300 items of data can be defined for each post-mortem.

The purpose of this project is to use data science analytic techniques to develop operational strategies that can be applied to paediatric post-mortems to prioritise which data is required to achieve the target of specifying the cause of death.

## Literary Review

Synopsis of project proposal but NOT a copy. Concentrate on parts that are directly relevant to the project.

## Report outline

Highlight some key aspects of the report, include some interesting graphics clearly annotated.

## Aim and Objectives

Take from project proposal; amended based on experience undertaking project?

The aims of this project are:

1. To develop a routine to extract data from the existing Post-mortem Research Database into an entity attribute value schema that will make the data more readily available for data analytics.
2. To apply the Decision Tree Analytical method to the extracted data to develop operational strategies that can be applied to paediatric post-mortems to prioritise which data is required to achieve the target of specifying the cause of death.
3. To investigate ensemble strategies, specifically Random Forests and Gradient Boosting to see how these techniques can improve on the basic Decision Tree method.
4. If time permits a final aim would be to consider if the approach of Neural Networks could be employed to enhance the results.

# Methods

Explain how the data was gathered and analysed

This section may be divided into sub-sections describing the conceptual design work and the actual implementation separately. Any problems or difficulties and the suggested solutions should be mentioned. Alternative solutions and their evaluation should also be included.

* 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.
* Project pipeline
  + PM Research Database => EAV Schema => Summary and Study based Attributes => RDV production => RDV Adjustments => Run models => review results
* Testing and Profiling

# Data

Explain why the data matches the objectives.

## Source

## Data Cleaning

* The nature of the data is that it is not normal
* Use of EAV schema to structure and clean data
* Introduce summary data for analytics and reporting

## Missing data

* Gestation in days at birth.
* Balanced data

## Normalisation vs Standardisation

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

# Analysis

Explain what analysis was undertaken and why.

## Create models

## Tune models

# Results

Describe the results of the analysis.

# Conclusion

## Project Summary

## Project Evaluation

## Recommendations for Future Work

# 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 – Example Project Code

Python Functions

Python ODBC

R Model

# Appendix B – Example RDV structure

Basic

Adjusted

# Appendix C – Detailed output from and analytic process

RDM file output

# Appendix D – Deliverables on Attached CD ROM

# 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.
* Balanced data
* Methods before data?
  + Include Software architecture in methods?
    - Tools and programing language
      * Why use R over Python?
    - Project pipeline
      * PM Research Database => EAV Schema => Summary and Study based Attributes => RDV production => RDV Adjustments => Run models => review results
  + Testing and Profiling
* Agile approach
  + Reviewed progress from last week
  + Set objectives for the coming week