

STUDENT & UNIT DETAILS – TO BE COMPLETED BY THE STUDENT

Candidate Number	03901
Unit Name and Code	ES50156 Practice Track

DECLARATION – PLEASE READ CAREFULLY

When you enrolled as a student at the University of Bath, you agreed to abide by the University's rules and regulations and agreed that you would access and read your programme handbook. This handbook contains references to, and penalties for, unfair practices such as collusion, plagiarism, fabrication or falsification. The University's Quality Assurance Code of Practice, [QA53 Examination and Assessment Offences](#), sets out the consequences of committing an offence and the penalties that might be applied.

By submitting this assessment, you confirm that:

1. You have not impersonated, or allowed yourself to be impersonated by, any person for the purposes of this assessment.
2. This assessment is your original work and no part of it has been copied from any other source except where due acknowledgement is made.
3. You have not previously submitted this work for any other unit/course.
4. You give permission for your assessment response to be reproduced, communicated, compared and archived for plagiarism detection, benchmarking or educational purposes.
5. You understand that plagiarism is the presentation of the work, idea or creation of another person as though it is your own. It is a form of cheating and is a very serious academic offence that may lead to disciplinary action.
6. No part of this assessment has been produced for, or communicated to, you by any other person.

MARK AND COMMENTS – TO BE COMPLETED BY THE MARKER

	MARK (%)

START YOUR ASSESSMENT HERE.

BEFORE YOUR ANSWERS, STATE EXPLICITLY WHICH PART / QUESTION / SUBQUESTION YOU ARE ATTEMPTING.

Department of Economics

Faculty of Humanities and Social Sciences

Practice Track Report for the Degree of MSc Economics for Business Intelligence and Systems

**Informed Application of Machine Learning
in predicting injury risk**

George Archer

October 2024

Faculty of Humanities and Social Sciences

Any student found to have cheated or plagiarised in assessment will be penalised. The Board of Examiners for Programmes will determine the nature and severity of the penalty but this may mean failure of the unit concerned or a part of the degree, with no provision for reassessment or retrieval of that failure. Proven cases of plagiarism or cheating can also lead to disciplinary proceedings as indicated in University Regulation 7.

I am aware of the guidelines on plagiarism: this coursework is the product of my own work.

Name: GEORGE ARCHER

Signed: 

Degree: MSC ECONOMICS FOR BUSINESS INTELLIGENCE AND SYSTEMS

Supervisor: DRAGO INDJIC

Word Count: 3995 **Date:** 30/09/2024

This report may be available for consultation within the University Library and may be photocopied or lent to other libraries for the purposes of consultation.

Copyright of this report rests with the author. No quotation from the report and no information derived from it may be published without the prior written permission of the author.

Informed Application of Machine Learning in Predicting Injury Risk

George Archer

MSc Economics for Business Intelligence and Systems (EBIS)

University of Bath

October 2024

Acknowledgements

First, I would like to extend my sincere thanks to Drago Indjic from Oxquant for his consistent and readily available support throughout this project. His guidance in helping me iron out a direction for the Practice Track once it became clear that real data would not appear has been invaluable.

I am also grateful to my family and friends, whose moral support has helped sustain my motivation and overcome challenges throughout the course of this project.

Abstract

Introduction: Recent smartphone and accessory innovation, i.e. smart watches and rings, has allowed high quality health data collection from individuals with ease.

Purpose: This project aims to provide a means to analyse user health data and, with machine learning, predict injury risk to identify users needing intervention.

Methods: Lacking a real-world dataset, promised by Optimi Health, the study was adapted to protect the Practice Track. It involved artificial data construction to mimic the intended structure and real instances, which involved combining available datasets, studies and assumed distributions. For privacy, and to enlarge the relational dataset, the Synthetic Data Vault library has been used to generate synthetic data in concert with ADASYN to address the common significant class imbalance. Meanwhile, the predictive ability of multiple models constructed using a range of learning methods: hyperparameter tuning, cost matrix, were analysed and compared.

Results: Using ADASYN, models were trained on 343 injury and 480 non-injury instances. The Python functionality enabled the creation of high-quality predictive models, supported by five-fold stratified cross-validation analysis. A decision tree model performed best, with $F2 = 0.581$ and correctly predicting 74.1% of injuries. However, it is evident there is need for greater sport scientist input in data creation with robust random forest and SVM being outperformed by validation models.

Conclusion: The study has developed means to generate robust decision tree, random forest and SVM models, which should prove invaluable when retrained on sufficient quality data. The artificial data creation provides a solid base which Optimi Health can improve on with sport specialists.

Keywords: Injury risk prediction; machine learning, modelling, synthetic data

Contents

1 Introduction	1
1.1 Aims	1
1.2 Background	2
1.3 Parameters.....	3
1.4 The project	4
2 Literature Review	5
2.1 Injury risk biomarkers	5
2.2 Synthetic data generation (SDG) methods.....	5
2.3 Machine learning models	7
3 Data	10
4 Methodology	12
4.1 Context:.....	12
4.2 Artificial data creation	14
4.2.1 Approach	14
4.2.2 Real data	14
4.2.3 Assumption-based data	14
4.2.4 Resultant data	15
4.3 Synthetic Data Generation (SDG).....	15
4.3.1 Synthetic data vault (SDV).....	15
4.4 Data preprocessing	17
4.4.1 Input	17
4.4.2 Time series aggregation	17
4.4.3 Model data.....	17
4.4.4 Sampling – ADASYN	18
4.5 The models	18
4.5.1 Brief	18
4.5.2 Baseline.....	20
4.5.3 OneR	20
4.5.4 Decision Trees.....	21

4.5.5 Random Forest.....	22
4.5.6 Support Vector Machine (SVM)	22
4.6 Model improvements	23
4.6.1 Aims.....	23
4.6.2 Cost Matrix	23
4.6.3 Hyperparameter tuning.....	23
4.7 Model evaluation	27
4.7.1 Traditional metrics	27
4.7.2 Stratified Cross-Validation (SCV).....	28
4.8 Implementation details	28
5 Artificial Data Visualisation	29
6 Results and Discussion	32
7 Next Steps.....	34
7.1 Data	34
7.2 Modelling techniques	34
7.3 Motion capture smartphone technology	35
8 Conclusion.....	36
9 Reference List	37
10 Appendix	57
10.1 Biomarker Literature Review	57
10.2 Synthetic Data Vault (SDV) Methodology	64
10.3 OneR Model Method	67
10.4 Model Evaluative Metrics	68
10.5 Decision Tree Model Graphic	71
10.6 Models with hyperparameter tuning using Recall scores	72
10.7 Glossary	75
10.8 ReadMe File	77

1 Introduction

This project, supervised by management consultancy Oxquant, is for start-up health company Optimi Health (OH), who aim to provide accessible health assessment to predict injury risk and, accordingly, prescribe personalised preventative exercise plans via mobile app.

There were clear challenges, specifically, my lack of sport science expertise and OH's lack of a clear research roadmap.

1.1 Aims

This paper initially aimed to contribute:

- 1) Machine learning (ML) models to predict injury risk – reviewing incumbent literature and using a provided real-world dataset.
- 2) A data pipeline from the app to the model

However, despite promises and reminders, no real-world dataset became available from OH. Once this deficiency became apparent halfway through, to mitigate risk to the Practice Track, I revised deliverables to establishing:

- 1) Artificial dataset
- 2) Synthetic data generation (SDG)
- 3) Predictive injury models

The study managed to cover all areas to high-quality: establishing functionality for random artificial dataset creation, review and application of appropriate SDG techniques, and analysis of multiple predictive models.

1.2 Background

Musculoskeletal injuries affect personnel in all occupations. However, injury likelihood through assessment of an individual's physical health is predictable.

Injury can severely dampen ability to perform physical activity and can result in permanent reductions in potential. Ekstrand et al. (2016) found that following ACL reconstruction surgery in professional football, only 65% of participants returned to pre-injury level. Pasqualini et al. (2022) find similar results following shoulder repair.

Beyond individual physical and mental wellbeing, injuries can greatly cost professional sports teams. Eliakim et al. (2020) estimated the financial effect on English Premier teams via season underperformance, calculating around 271 cumulative days out for injury may result in falling one place. In 2016-17 season teams on average endured 1410 such days – an average six-place drop versus expected finish, equating an estimated £36m cost. Hence, injury mitigation ought to be prioritised, especially given injured player wages averaging £9m, culminating in an average injury cost of £45m per season.

Meanwhile, given athletes generally have short times at peak performance, injury can be career changing. Loberg's (2009) review finds injury a principal reason for involuntary career termination. Batt et al. (2021) found 21.8% of survey respondents (retired British Olympians) reported retiring early due to injury.

1.3 Parameters

Targeted for real world use, data collection constraints exist due to equipment and technical requirements. Hence, the models involve collectable inputs which can be accurately recorded without trained professionals.

GDPR legislation requires clarity on the purpose of collected data and UK Government(2017) states “*The regulation of health apps provides patients and healthcare professionals the assurance that apps are high quality, safe and ethical.*”.

These concerns are warranted by numerous health-related apps lacking quality and adherence to clinical recommendation – Hajratalli et al. (2019) examined 120 top health apps finding around 50% made near medical claims and 35 to raise “serious concerns regarding safety”. Further highlighted when three of four apps incorrectly classified at least 30% of melanomas as un concerning (Akilov et al.,2013).

OH have requested limits on scope so its app is not recognised as providing clinical diagnosis – likely to ensure labelling as a fitness application involved in prevention and not a medical device (Schmitz,2022). However, given clients include the British Army, and sport leagues, the model requires sufficient academic backing.

1.4 The project

Multiple ML models are trained and compared for injury risk prediction, including baseline, custom OneR, decision tree, random forest and support vector machine algorithms. To address large class imbalance, ADASYN increases injury instances to help train the models, which are then optimised via hyperparameter tuning to personalise them to data. Stratified cross-validation and focus on recall and F2 scores provide extensive assessment of predictive power placing emphasis on the correct prediction of injury instances. The decision tree performed best, correctly predicting 74.1% of injuries. However, lacking data necessarily reflective of real instances, other robust complex models were outperformed by OneR on test data.

2 Literature Review

This section provides academic context on three critical elements.

2.1 Injury risk biomarkers

Injury is widely recognised as a multi-facet event, with interacting and confounding effects making it difficult to ascertain aetiology (Meeuwisse,1994; Bahr and Holme,2003).

Researchers commonly analyse a broad selection of variables to predict injury risk, often including personal statistics, injury history, wellbeing, training load, body composition, equipment, motion, biomechanics, wearable data and more (Ayala et al.,2018; Cintia et al.,2021)

To sustain data science focus, the complete biomarker literature review is in [Appendix 10.1](#).

2.2 Synthetic data generation (SDG) methods

To accurately predict, models require data in sufficient quantities and quality (Birişçi et al.,2022). Synthetic data provides an alternative when real data is unavailable, sparse, private, or for testing.

Its use applies here where OH, in seed capital stage, lack a dataset for predictive modelling. Generation of synthetic tabular data represents an invaluable asset to constructing a predictive model which can then predict user injury risk till sufficient real data surfaces to re-evaluate and retrain.

Androutsos et al. (2024) define various SDG models and provide an extensive review of incumbent synthetic tabular data methods including Python libraries used. Deep-learning models are prime models, the likes of GAN and VAE popular.

Generative adversarial networks (GANs) are composed of three aspects: real data sample, and two neural networks – a generator to create synthetic data and the discriminator which attempts to distinguish the real from synthetic (Chen et al.,2020; Bottarelli et al.,2022). If the synthetic data fails to “deceive” the discriminator it is generated until it does, culminating in highly realistic synthetic data.

Variational auto encoders (VAEs) comprise of two components: encoder and decoder. The encoder maps an input vector (real data point) to a latent vector in a latent, lower-dimensional space (Chandrasekaran et al.,2023). The decoder takes this latent vector as an input and tries to construct the original input vector (real data point) and optimise the reconstruction error. In layman’s terms, VAE learns the distribution of data and from this creates synthetic data. Both deep-learning approaches are found to provide accurate representations of real instances.

The synthetic data vault (SDV) python library offers many SDG models, including GAN and statistical-based Gaussian Copula (Androutsos et al.,2024; Datacebo,2023a). Its key impressive feature is its malleability for handling multiple dataset types, specifically relational databases. The “Multi-Table Data” synthesisers provide statistical-based learning via Gaussian Copula processes to attain joint distributions and generate data (Datacebo,2023b). Experimentation proved it provides viable synthesised data with no statistically significant difference (Patki et al.,2016).

Without real data to create synthetic data, statistical information and stylised facts can be used to generate data (Datacebo,2023b;Sciki-learn,2024e).

2.3 Machine learning models

In their review of injury prediction ML methods, Van Eetvelde et al. (2021) confirmed their validity and usefulness for decision-making for prevention and prediction for trainers and medical practitioners. They highlight diversity in popular modelling techniques: tree-based, support vector machines (SVMs), neural networks, random forest and ensemble methods. For injury prediction a classification model is desirable. Amendolara et al. (2023) and Kumar et al. (2024) summarise ML techniques used for injury prediction where a trade-off between transparency and simplicity versus higher accuracy and computational resources is apparent.

Multiple studies utilise decision trees for their interpretability and ability to handle both numerical and categorical data and have found them to provide reasonable predictive power (Amendolara et al.,2023). They define a set of conditions in a hierarchical structure. An instance can be classified by following the path dictated by satisfied conditions till it reaches a leaf node – stating its classification.

Blumkaitis et al. (2023) found an SVM to yield the best results for injury prediction, with 96.3% of data points correctly detected. SVMs are recognised as powerful tools using mathematical model to manipulate data such that a division of the domain is possible – defining classes and enabling classification prediction (Huang et al.,2018). This is effective given its suitability for predicting high-dimensionality datasets and the complex multi-facet nature of injury (Amendolara et al.,2023). However, it is computationally expensive and, given prospective dataset size, may place limits on use.

Random forest algorithms are commonly used and proven to show comparative promise for injury prediction (Ruddy et al.,2018). Farhadian et al.'s(2020) prediction

of sports-related dental injuries found random forest to have 89.3% accuracy suggesting applicability.

Amendolara et al. (2023) cover a host of neural network techniques, highlighting that they “*tend to be the most accurate and powerful*”. They specifically mention Convolutional (CNNs) and Recurrent Neural Networks (RNNs) as highly favoured. Yet, note the high performance comes with increased complexity, computational requirements and training time.

Kautz et al. (2017) found CNNs to provide greater accuracy than other algorithms of 83.2% - exceeding others including SVM and decision tree by 16% when monitoring beach volleyball players. Ghazi et al. (2021) attained a prediction accuracy of concussion cases over 90%, making CNN attractive for quality purposes.

Meng and Qiao (2021) explored the use of multiple neural networks, and combined CNN with long short-term memory (LSTM) attaining 97.54% specificity and 97% accuracy - suggesting CNNs may be best for client-use.

Dhanke et al. (2022) compare RNN to SVM when predicting the effect of physical training on injury likelihood and found it outperforms with 99% accuracy. A RNN may provide additional predictive power due to its design to capture temporal correlations compared to other algorithms. Given injury prediction is reliant on changes to physical health, this additional temporal consideration is valuable.

Several academics explore more contextual approaches, accounting for the recurrent and residual nature of injuries (Ullah et al.,2012;Zumeta-Olaskoaga et al.,2021; Hägglund et al.,2006;De Visser et al.,2012). Alongside Bikandi et al.(2023), they explore recurrent time-to-event modelling for injury prediction with shared frailty Cox models.

Cox proportional hazard models are part of the survival analysis field and define the hazard function (age-specific failure rate) as a function of covariates and unknown regression coefficients multiplied by an unknown time function, providing estimation of relative injury risk (Cox,1972). The shared frailty term accounts for unobserved heterogeneity such as individual-specific factors including different genetic traits that affect injury likelihood. It introduces a random effect allowing the model to handle dependence between recurrent events (injuries) within the same individual.

3 Data

The artificial data is based off expected data structure and metadata. The synthetic dataset is constructed via a combination of real-world datasets (to retain relations) and assumed distributions for harder to source data.

The user-inputted database from OH is scheduled to comprise of six dataframes, defined in Figure 3.2, all connected by user id per Figure 3.1.

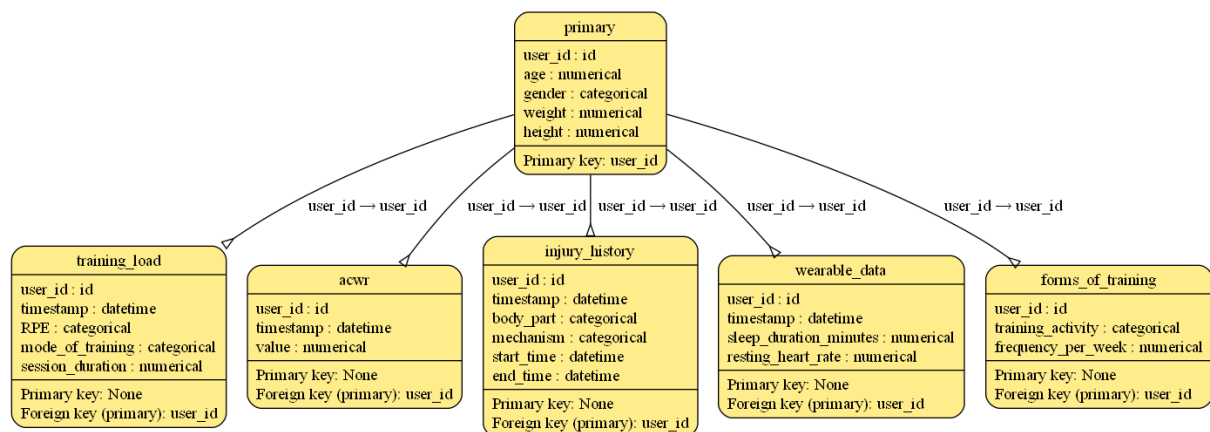


Figure 3.1 Injury risk dataset metadata

See [4.2](#) for details on artificial dataset creation.

Figure 3.2: Dataset details, by dataframe

Dataframe	Variables	Defintion
Key	User ID	User identification
Primary	Age	Years lived
	Gender	Interpreted as sex so variable is binary, where female is 0 and male is 1
	Height	In centimetres
	Weight	In kilograms
Training Load	Timestamp	Date of performed activity
	Rate of Perceived Exertion (RPE)	Subjective measure on a 0-10 scale
	Mode of Training	Defines form of activity, i.e. Basketball
	Duration	Time of session in minutes
Acute Chronic Workload Ratio (ACWR)	Timestamp	Date of performed activity
	ACWR Value	Ratio assessing relative personal change in activity*
Injury History	Timestamp	-
	Body part	Represent general injured body part
	Mechanism	Describes inciting event setting of injury
	Start	Date injury occurred
	End	Date deemed to have recovered
Wearable Data	Timestamp	
	Sleep duration	Time slept, minutes
	Resting Heart Rate	Beats per minute when at rest
Forms of Training (Contains multiple for same user)	Timestamp	-
	Training Activity	Defines form of activity, i.e. Basketball
	Frequency per week	Defines number of times <i>Training Activity</i> performed per week

*see glossary for definitions

Figure 3.2 Dataset details, by dataframe

4 Methodology

4.1 Context:

With the promised dataset absent, and as suggested Drago Indjic (Oxquant), this study has had to expand to develop three tools:

- 1) artificial dataset creation
- 2) application of a relevant synthetic data generator
- 3) consequent predictive modelling for injury

Kim and Malikov (2022) recognised Python as the preferred language for data science.

All Python programs developed are available on [Github](https://github.com/aca247/Injury_Risk_Prediction.git):

https://github.com/aca247/Injury_Risk_Prediction.git

Figure 4.1 summarises the constituent parts.

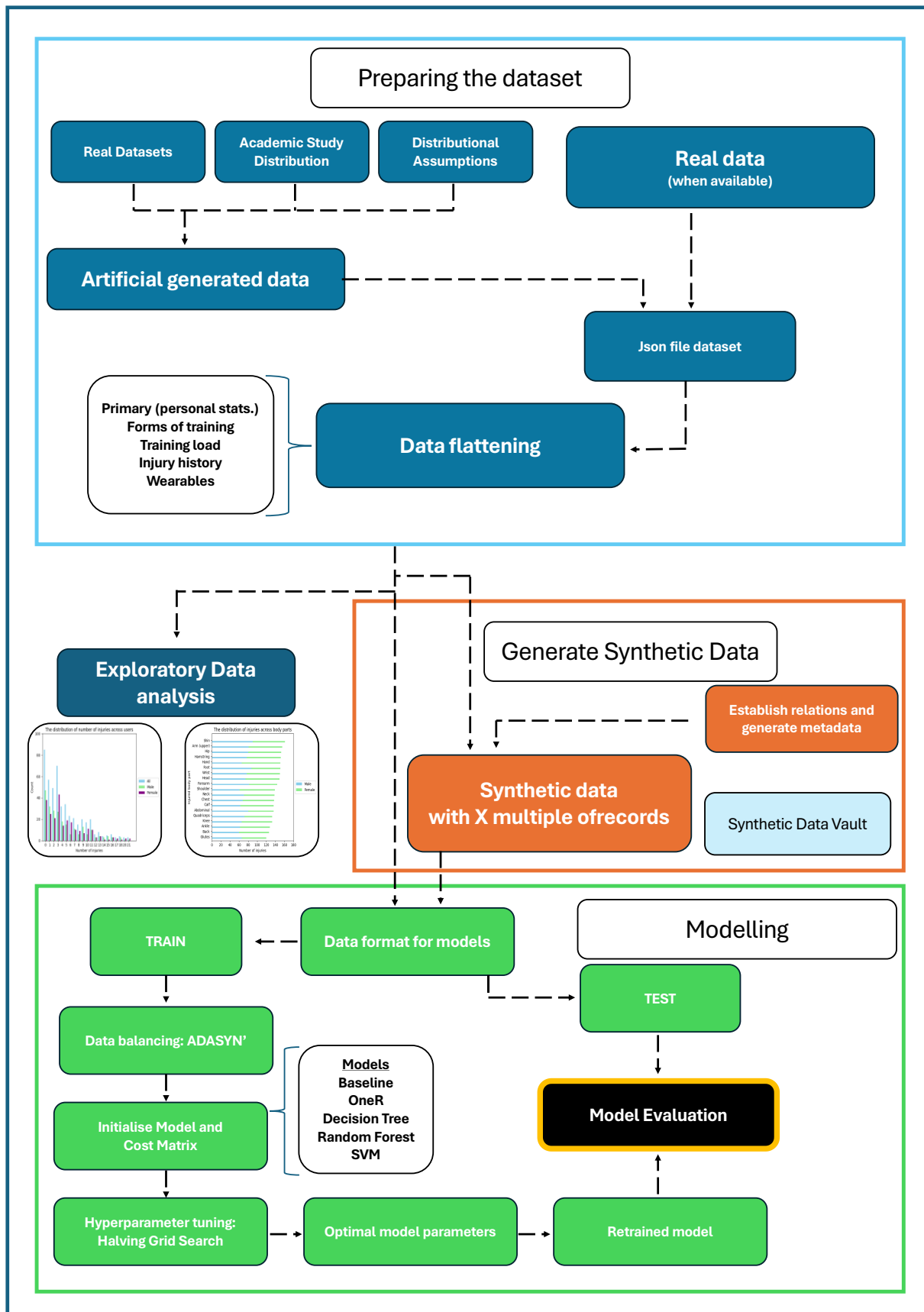


Figure 4.1 Data pipeline schematic

4.2 Artificial data creation

4.2.1 Approach

To create artificial data, a hybrid approach combines real data and rule-based statistical modelling. This is because it is difficult and time-consuming to find data reflecting real instances for all variables. This is exacerbated by the likely relational nature of real data when creating realistic representation – hence its use where possible.

4.2.2 Real data

Incumbent datasets and studies garner some insight. The NHANES 2017-18 survey, from the National Center for Health Statistics (NCHS)(2018b;2020a;2020b), provides personal statistics. Meanwhile, distributions for sleep and resting heart rates (RHR) are attained from incumbent studies (NCHS,2018a;Galarnyk et al.,2020). The relatively static nature of RHR allows generation to randomly oscillate around steady state values to produce a time series, while sleep duration randomly drawn from a distribution (Galarnyk et al.,2020).

4.2.3 Assumption-based data

Other variables, injury history and training habits, are less available. Hence, it is assumed that users are predisposed to active lifestyles – reducing need to account for sedentary behaviours (Bond et al.,2017).

Aizen's Theory of Planned Behaviour (1991) predicts that planned behaviours are determined by intentions. Given users have divested interest in physical health, their behaviour may be predictable, i.e. weekly classes (Biddle et al.,2002; Jiang and Mengru,2022). Hence, historical training records can be generated by assigning

routine according to randomly endowed user types. This entails randomly generating session durations and RPE from generic left-skewed distributions to provide routine records since signup with randomness introducing cancellations.

Injury history is difficult due to complex injury aetiology. Hence, without overfitting data and an unnecessarily complex algorithm, injury history is randomly assigned over lifespans following cumulative injuries-to-date endowment from a right-skew distribution, see [Figure 5a](#).

4.2.4 Resultant data

Real data coupled with assumed distributions creates artificial user records to resemble the intended JSON dataset. Functionality allows different random generations of artificial datasets.

However, the predictive inference from this data should be treated guardedly with it potentially not sufficiently reflecting randomness and thus allowing for interpretable patterns by the ML algorithms.

In contrast, the generated data may not capture fundamental aspects of injury aetiology and thus make it difficult for algorithms trained on it to correctly predict injury.

4.3 Synthetic Data Generation (SDG)

4.3.1 Synthetic data vault (SDV)

Given multivariate effects, an SDG to handle relational data was important. Most look at single table generation rather than multi-table, limiting use of GANs and VAEs (Chen et al.,2020).

The Synthetic Data Vault provides a full means through its conditional parameter aggregation (CPA) which specifies how child and parent tables relate. (Patki et

al.,2016). The free HMASynthesiser requires two inputs: original dataset and metadata ([example](#)).

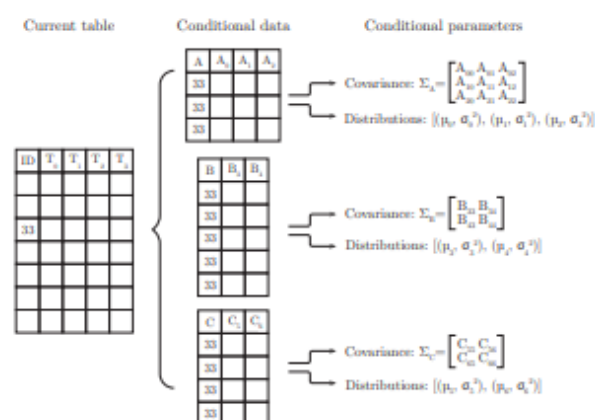


Figure 4.3.1a Example of CPA for row iteration with key “33” (Patki et al.,2016)

The CPA iterates through each row (*user_id*) of the parent, generating conditional data for each child containing the *user_id* and then calculating joint distributions via the Gaussian Copula process, as Figure 4.3.1a.

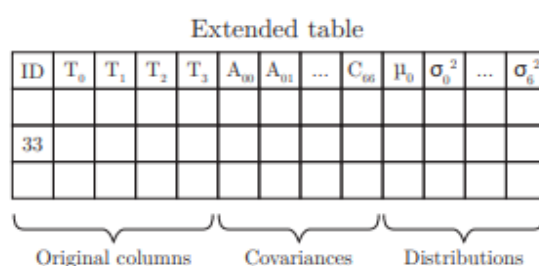


Figure 4.3.1b: Result of CPA (Patki et al.,2016).

This creates an extended table, from which the Gaussian Copula process captures cdfs and covariances which can generate a synthetic dataset – more in [Appendix 10.2](#).

4.4 Data preprocessing

4.4.1 Input

A deep data flattening approach is required to reformat data a complex-nested dictionary with spurious records for the SDG, EDA and predictive modelling.

The artificial dataset is not designed for missing data. Approaching missing values, imputation is unsuitable due to the relational nature which could lead to incorrect diagnosis – or, given injury-prone data minority, create bias toward no-injury predictions. Instead, separate models could be trained on constrained datasets and the app return a prompt: *“for more accurate predictions, input x”*

4.4.2 Time series aggregation

With high day-to-day volatility of variables possible, feasibly resultant of external factors, i.e. stress, raw daily analysis is likely not interpretable for injury risk. Coyne et al. (2022) provide support for smoothing methods. Hence, the daily dataset is converted to weekly for model training and testing – with training forms, i.e. basketball, hot encoded and filled with respective week duration values. Simple moving averages represent desirable future adaptation.

4.4.3 Model data

With few positive instances in data snapshots, for snapshot analysis, weeks are assumed independent, as users and their activity are. This means any arbitrary collection of weekly user records could form a dataset, i.e. week of 09/01/2017 for one user and 26/08/2024 for another, and train the model.

This promotes training and test data that contain sufficient positive instances. This is needed for predictive modelling, and data resampling techniques require a minimum to generate additional positive instances.

4.4.4 Sampling – ADASYN

Injury risk datasets suffer from significant class imbalance with few positive (injury) cases relative to negative (no-injury) cases. Cintia et al. (2021) found only 2% of instances involved injury, making training of models difficult.

Over- and under- sampling provide means to adjust class distribution, with over-increasing minority instances, while under- reduces majority instances. Under-sampling results in possible information loss making it less attractive.

With a highly imbalanced dataset, over-sampling technique ADASYN is applied to the training set, using Python library *imblearn* (Cintia et al.,2021;Imbalance-learn developers,2024;Bai et al.,2008).

4.5 The models

4.5.1 Brief

Ayala et al. (2018) utilised four decision tree algorithms to attain founded conclusions. Similarly, various models were conducted and their performance assessed for injury prediction. Figure 4.5.1a outlines the models and necessary python libraries while Figure 4.5.1b outlines tools to assess and improve model predictive power.

Figure 4.5.1a: Machine Learning Models*		
Model	Python dependencies	R dependencies
Baseline	User identification	
Custom OneR	sklearn.tree • DecisionTreeClassifier	
Decision Tree	sklearn.tree • DecisionTreeClassifier	
Random Forest	sklearn.ensemble • RandomForestClassifier	
Support Vector Machine (SVM)	sklearn.svm • SVC	
OneR	rpy2	OneR
Shared Frailty Cox Model	rpy2	survival

*Only covers primary dependencies, not other workings

Figure 4.5.1a Machine Learning Models and dependencies

Figure 4.5.1b: Machine Learning Models other tools*		
Tool	Python dependencies	R dependencies
ADASYN	imbalanced_learn • ADASYN	
Confusion Matrix	sklearn.metrics	
Stratified Cross Validation (SCV)	sklearn.model_selection • cross_val_score • cross_val_predict • StratifiedKFold	
GridSearch	sklearn.experimental • enable_halving_search_cv sklearn.model_selection • HalvingGridSearchCV	
Synthetic Data Generator	sdv_metadata_generator sdv.multi_table • HMASynthesiser	

*Only covers primary dependencies, not other workings

Figure 4.5.1b Machine Learning Models other tools and dependencies

Unfortunately, common complication with library rpy2 versions – allowing R functionality in Python – resulted in a custom OneR and shared frailty Cox models not constructed.

4.5.2 Baseline

For model validation a baseline is used. If it outperforms ML models, they should not be considered valid.

For simplicity, scikit-learn DummyClassifier is used. This predicts positive instances via a stratified approach – model trained to randomly predict x% of instances as positive because the training data contained x% positive instances.

4.5.3 OneR

To further validate complex methods, the OneR model provides simple classification.

It resembles a decision tree restricted to one level, where one rule that attains the lowest predictive error defines classification (Von Jouanne-Diedrich,2017). Sayad (2024) explains in [Appendix 10.3](#).

This provides simplicity and, in some studies, outperformed complex models (Singh,2017). The custom algorithm used the DecisionTreeClassifier combined with grid search.

The prime difference being my approach optimally chooses Entropy or Gini criterion while traditional OneR minimises classification error to decide the rule.

4.5.4 Decision Trees

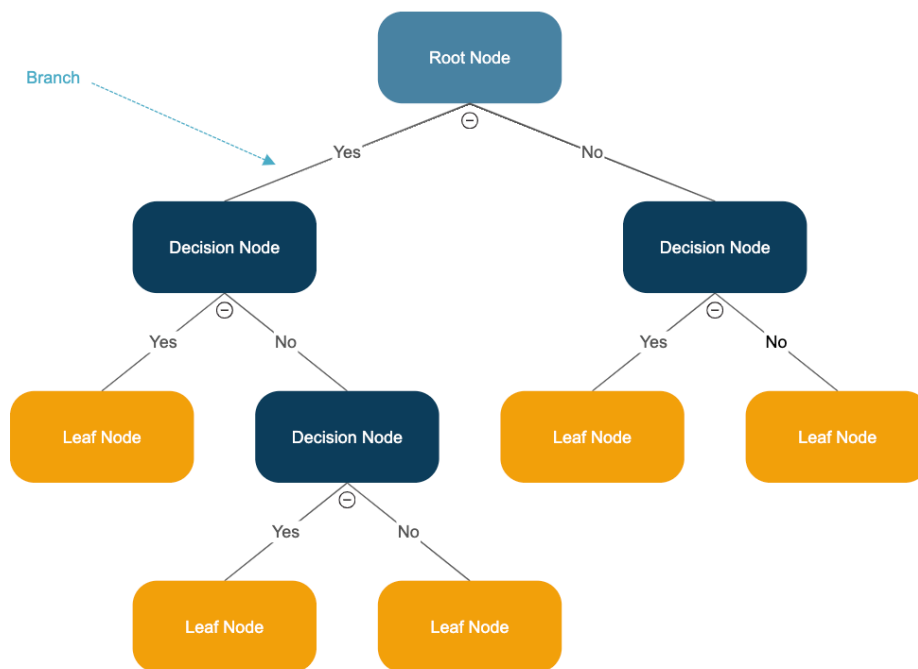


Figure 4.5.4 Decision Tree Structure (Smartdraw,2024)

Decision trees can provide clear interpretation of injury aetiology via hierarchical structure and risk classification according to chain of satisfied conditions, as Figure 4.5.4. They account for the multi-facet nature rather than just univariate effects.

Scikit-learn(2024c) offers `DecisionTreeClassifier` which represents an optimised version of CART. In Python, CART is more accessible than alternatives C4.5 /C5.0, available in R, and requires less computational resources allowing faster realisation.

4.5.5 Random Forest

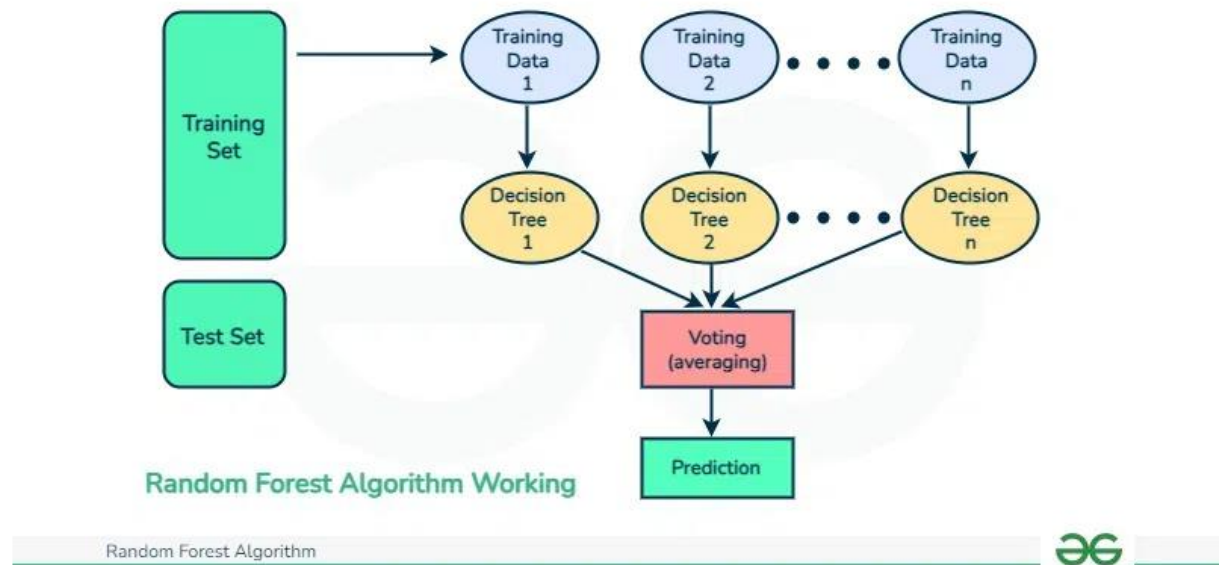


Figure 4.5.5 Random Forest (GeeksforGeeks,2024b)

Random forests represent an ensemble method involving a collection of decision tree predictors where the constructed trees vote on the class for a given inputted instance – the most popular specification represents the prediction (Breiman,2001). Figure 4.5.5 illustrates the process of training multiple decision trees using training data samples and how voting defines prediction.

4.5.6 Support Vector Machine (SVM)

SVMs classify individual instances by defining hyperplanes which best separate and classify data points in the n-feature space (Kumar et al.,2024). The numerous features of the dataset and performance in other studies make SVMs attractive.

Complexity of injury means that knowing the appropriate SVM kernel to define hyperplanes is difficult. To ensure its performance, hyperparameters are tuned according to [section 4.6.3](#), testing hyperplane construction in alternate dimensional spaces (Scikit-learn,2024b)

4.6 Model improvements

4.6.1 Aims

The project aims to maximise injury instances correctly predicted – maximising recall score. It is less concerning if users without imminent injury are mis-identified and prescribed preventative plans.

4.6.2 Cost Matrix

To prioritise correct classification of injury instances, a cost matrix places a greater cost on mis-specifying injury imminent instances. Figure 4.6.2 illustrates an example (models use sklearn ‘balanced’ weighting):

Cost Matrix		Predicted Class	
		No Injury Imminent	Injury Imminent
Actual Class	No Injury Imminent	TN Cost = 0	FP Cost = 2
	Injury Imminent	FN Cost = 6	TP Cost = 0

Figure 4.6.2: Example cost matrix

4.6.3 Hyperparameter tuning

Cintia et al. (2021) exaggerate the importance of hyperparameter tuning to control the behaviour of ML models and improve prediction – despite only finding two papers implementing the method. The grid search method exhaustively considers all hyperparameter combinations. Alternatively, the random search algorithm samples a number of combinations from a distribution of possible values, at benefit of lower

computational resources. Given the models should rarely require retraining, the grid search method seemed more appropriate.

Scikit-learn(2024a) offers GridSearchCV for this. However, accounting for computational resources and feasible dataset size, provides alternative HalvingGridSearchCV – an experimental function that performs similarly but uses less resources. Instead of searching all combinations, it searches the parameter space using successive halving – iterative selection process where all combinations are considered using few resources at first iteration (Scikit-learn,2024f). A subset of likely higher-scoring combinations, determined via F2-score, is processed in subsequent iteration with more resources, which continues till the best subset is found. This may not produce as good model, but significantly reduces runtime. To ensure optimal parameter combinations, the algorithm uses five-fold stratified cross-validation to find and test parameters. This removes possible train-test split bias by ensuring the model is evaluated on different data subsets – giving a more realistic assessment of model performance - see Figures 4.6.3a and 4.6.3b.

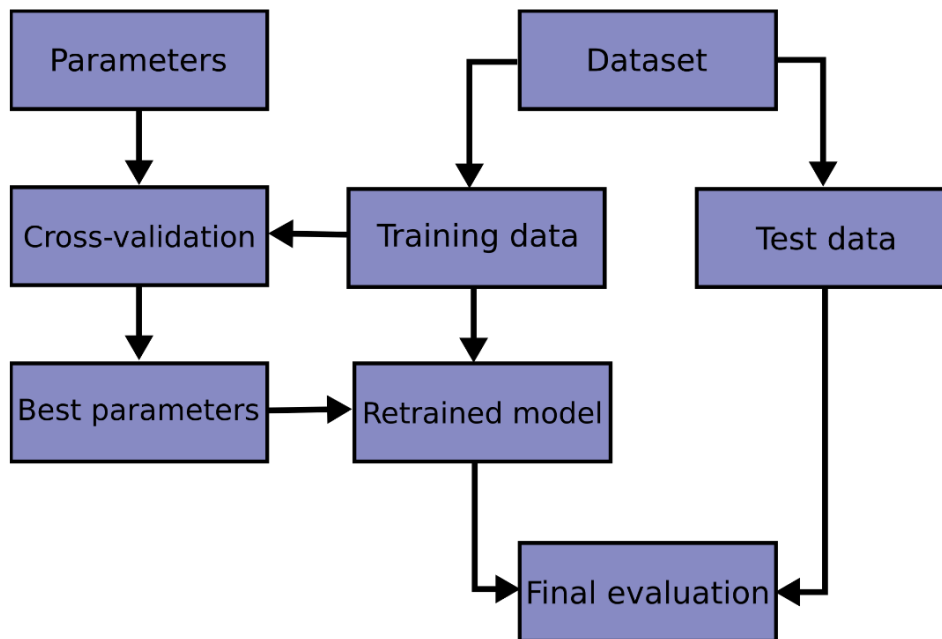


Figure 4.6.3a: ML model training with hyperparameter tuning (Scikit-learn, 2024d).

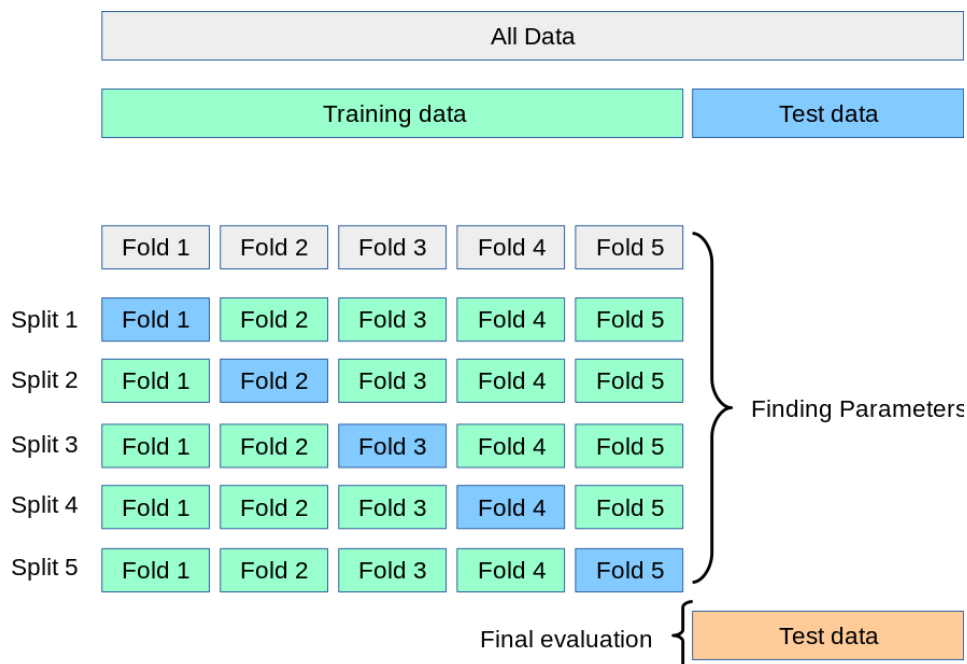


Figure 4.6.3b: Example of five-fold stratified cross-validation within hyperparameter tuning (Scikit-learn, 2024d).

Figure 4.6.3c lists the tuning process for each model with the fourth column describing optimal hyperparameters for models in [section 6](#).

Figure 4.6.3c: Machine Learning Models — Hyperparameter Tuning with F2-Score			
Model	Parameters to tune	Parameter grid set	Optimal parameters
Baseline	—	—	—
OneR custom	Minimum samples to split at decision node	min_samples_split : [2, 4, 6, 8, 10]	8
	Minimum samples at each leaf node	min_samples_leaf : [1, 2, 3, 4, 5]	5
	Function to measure split quality	criterion : ['gini', 'entropy']	Entropy
Decision Tree	Maximum number of decision nodes	max_depth : [3, 5, 7, 10, 12, 20]	7
	Minimum samples to split at decision node	min_samples_split : [2, 4, 6, 8, 10]	2
	Minimum samples at each leaf node	min_samples_leaf : [1, 2, 3, 4, 5]	1
	Function to measure split quality	criterion : ['gini', 'entropy']	Entropy
Random Forest	Number of decision trees constructed to vote on classification	n_estimators : [10, 20, 30, 50, 60, 75, 100]	100
	Maximum number of decision nodes per tree	max_depth : [3, 5, 7, 10, 12, 20]	12
	Minimum samples to split at decision node	min_samples_split : [2, 4, 6, 8, 10]	2
	Minimum samples at each leaf node	min_samples_leaf : [1, 2, 3, 4, 5]	1
	Random sampling with or without replacement when constructing decision trees	bootstrap : [True, False]	True
Support Vector Machine (SVM)	Regularisation parameter controlling aim to classify all points correctly at cost of more complex hyperplanes	C : [0.1, 0.5, 1, 1.5, 2]	0.1
	Degree of influence a single training instance has	gamma : ['scale', 'auto']	auto
	Defines the degree of the 'poly' kernel	degree : [2, 3, 4, 5]	2
	Defines the dimensional space and learning of the hyperplanes	kernel : ['linear', 'poly', 'rbf', 'sigmoid']	poly

Figure 4.6.3c: Details of hyperparameter tuning settings, including parameter grids for each model

4.7 Model evaluation

Multiple evaluative methods are used to ensure assess performance. As [4.6.1](#), the aim is to predict as many injuries as possible.

4.7.1 Traditional metrics

To gauge predictive power several metrics are used (Naidu et al.,2023):

- Recall and specificity
- Precision and negative predicted value
- F1-score
- F2-score
- AUC-ROC

Accuracy scores, although common, are not intuitive and can be misleading with imbalanced datasets. Metrics of interest are recall and F2 scores.

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

Recall tells actual injury instances correctly predicted. As the cost of misclassifying injury instances, and not prescribing treatment, is greater a higher recall is important.

The F2-score values the recall more than the precision score (Senbel et al.,2022):

$$F2 = \frac{(1 + 2^2) \times Precision \times Recall}{2^2 \times Precision + Recall}$$

This balances the greater need to predict injuries while aiming to minimise wasteful plan creation via false positives. Other metrics detailed in [Appendix 10.4](#).

4.7.2 Stratified Cross-Validation (SCV)

As [4.6.3](#), SCV provides a thorough means to evaluate models and their predictive power (Ayala et al.,2018). It removes reliance on a single train-test split whereby the random split may not be representative. SCV splits the training data into n folds with class distribution preserved. Figure 4.6.3b demonstrates how it trains the model on n-1 folds to then test on the remaining fold iteratively, till all combinations analysed. It then averages evaluation metrics to provide a model assessment.

4.8 Implementation details

To see the full set of programs, see [Github folder](#) or readme in [Appendix 10.8](#).

5 Artificial Data Visualisation

Having created artificial data, Figures below demonstrate resultant data snippets.

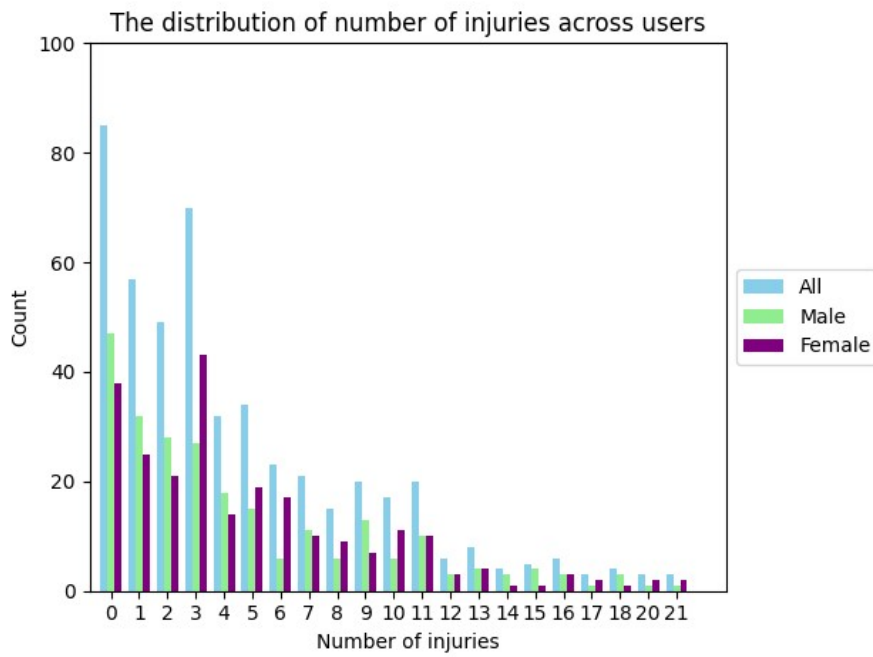


Figure 5a: Distribution of lifetime injuries, sex split

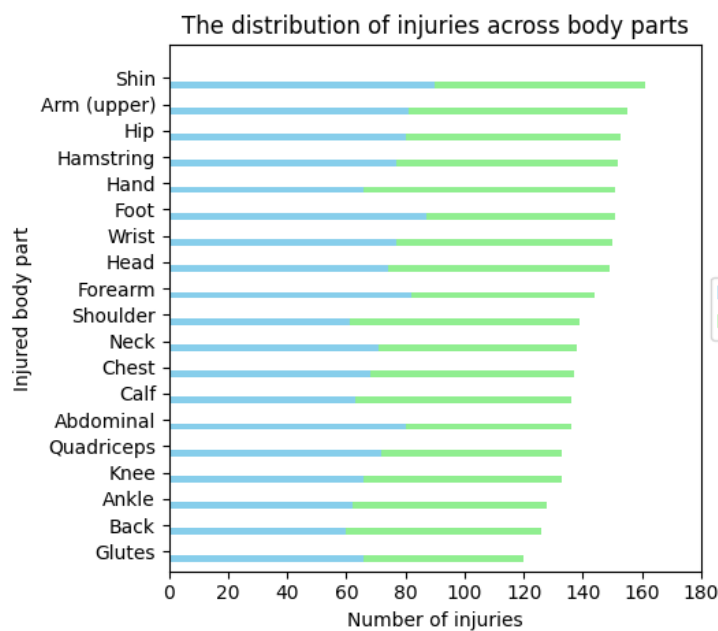


Figure 5b: Distribution of lifetime injuries across injured body parts

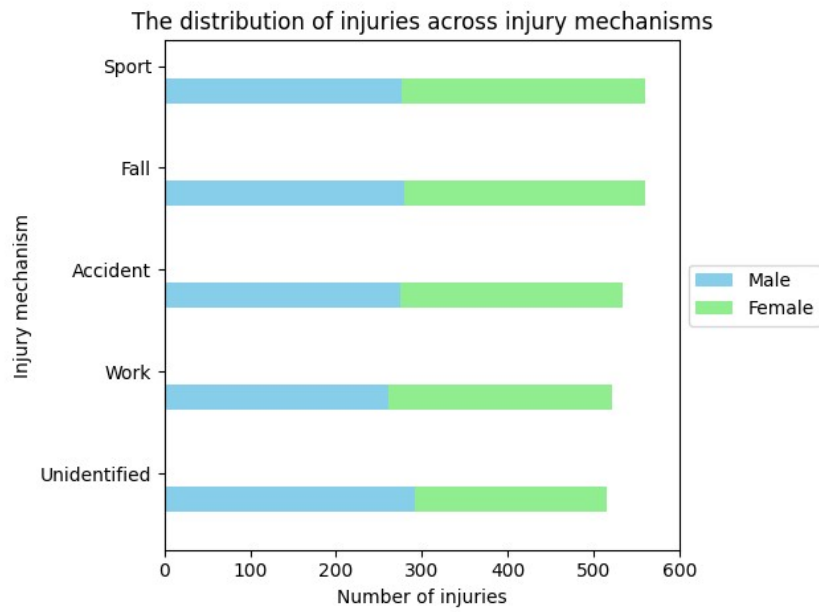


Figure 5c: Distribution of lifetime injuries across injury mechanisms

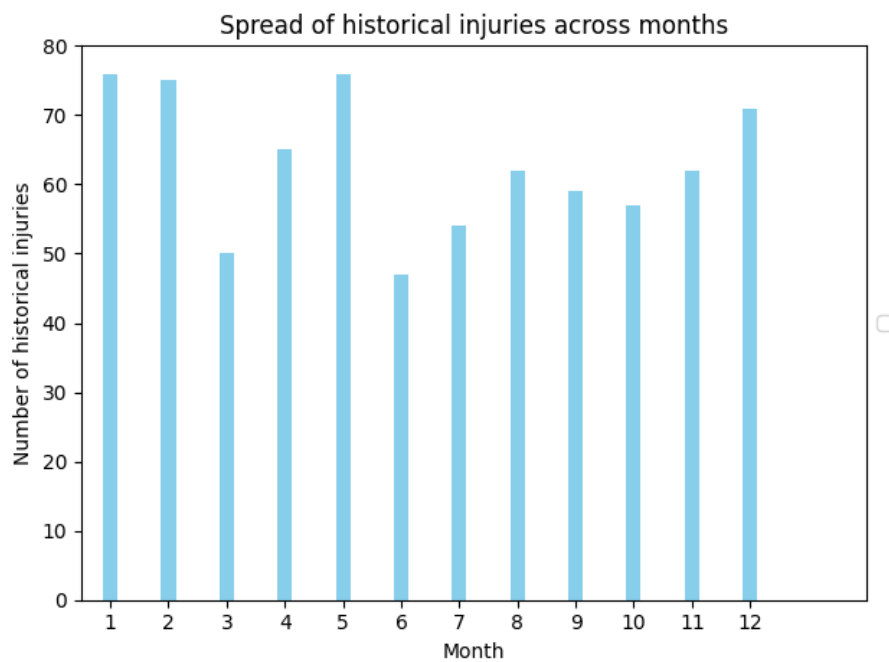


Figure 5d: Distribution of injuries across months of the year

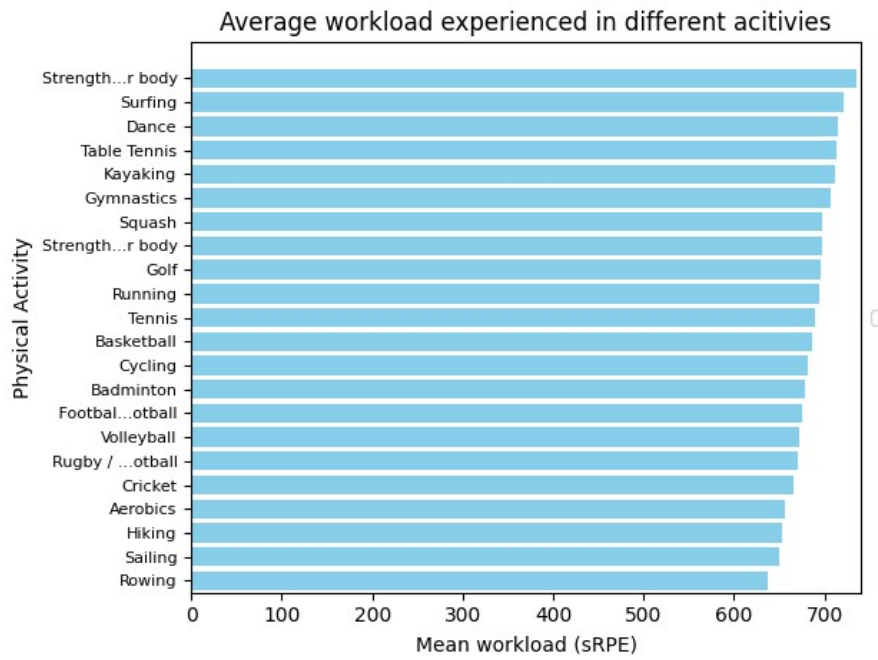


Figure 5e: Average workload (sRPE) per physical activity

6 Results and Discussion

The models used a 70:30 training-test split for the 650-user dataset. Following ADASYN on training data only, to avoid overfitting, models were trained on 343 injury and 480 non-injury instances using grid search techniques to [tune hyperparameters with F2-scores](#)— prioritising injury prediction while acknowledging non-injuries. Five-fold SCV evaluates the predictive performance of the different algorithms in Figure 6.1.

Figure 6.1: Full evaluative metrics for models, tuned on F2-scores — five-fold stratified cross-validation (ADASYN)								
Model	Class	Recall	Precision	F1-score	F2-score	AUC-ROC	Predicted	Misspecified
Baseline	No injury imminent	0.493	0.496	0.494	0.493	0.495	166	171
	Injury imminent	0.497	0.494	0.496	0.496	0.495	167	169
Custom OneR	No injury imminent	0.988	0.653	0.786	0.896	0.684	333	4
	Injury imminent	0.473	0.975	0.637	0.528	0.684	159	177
Decision Tree	No injury imminent	0.914	0.933	0.924	0.918	0.935	308	29
	Injury imminent	0.935	0.915	0.925	0.931	0.935	314	22
Random Forest	No injury imminent	0.985	0.979	0.982	0.984	0.999	332	5
	Injury imminent	0.979	0.985	0.982	0.980	0.999	329	7
Support Vector Machine	No injury imminent	0.893	1.000	0.944	0.913	0.953	301	36
	Injury imminent	1.000	0.903	0.949	0.979	0.953	336	0

*Sum Predicted and Misspecified columns represent total number of that class

Figure 6.1: Full evaluative metrics to assess predictive power of models using five-fold stratified cross-validation

Figure 6.1 demonstrates that complex models, decision tree, random forest and SVM algorithms are particularly robust for injury prediction with high recall and F2-scores.

Figure 6.2: Full evaluative metrics assessing performance of models, tuned on F2-scores, on test data								
Model	Class	Recall	Precision	F1-score	F2-score	AUC-ROC	Predicted	Misspecified
Baseline	No injury imminent	0.448	0.941	0.607	0.500	0.438	64	79
	Injury imminent	0.429	0.037	0.067	0.136	0.438	3	4
Custom OneR	No injury imminent	1.000	0.973	0.986	0.994	0.714	143	0
	Injury imminent	0.429	1.000	0.600	0.484	0.714	3	4
Decision Tree	No injury imminent	0.930	0.985	0.957	0.941	0.821	133	10
	Injury imminent	0.714	0.333	0.455	0.581	0.821	5	2
Random Forest	No injury imminent	0.993	0.959	0.976	0.986	0.793	142	1
	Injury imminent	0.143	0.500	0.222	0.167	0.793	1	6
Support Vector Machine	No injury imminent	0.860	0.953	0.904	0.877	0.428	123	20
	Injury imminent	0.143	0.048	0.071	0.102	0.428	1	6

*Sum Predicted and Misspecified columns represent total number of that class

Figure 6.2: Full evaluative metrics to assess predictive power of models using five-fold stratified cross-validation

However, Figure 6.2 shows random forest and SVM performed poorly on test data with $F2 \leq 0.2$. Validation models outperformed them with OneR scoring $F2 = 0.484$. Yet, the decision tree performed moderately well, returning $F2 = 0.581$ and 71.4% correct prediction of injury instances.

As per [section 4.2.4](#), artificial data generation is the likely cause of sub-optimal performance. Pattern identification and injury aetiology diagnosis is unlikely with artificial data not informed by a professional sport scientist or biologist – hence, not necessarily representing injury features.

Yet, the developed Python functionality allows dynamic high-level predictive model development which can be retrained on better quality data, when available. The interpretable models established allow clear understanding, as per decision tree in [Appendix 10.5](#), and random forest graphics in [GitHub](#).

7 Next Steps

This marks the first attempt predicting injury risk for OH. Possible improvements to both data and models exist.

7.1 Data

The biomarker literature review in [Appendix 10.1](#) describes numerous alternate attainable data measures. OH could explore providing high-level assessments as paid add-ons to provide deeper analysis.

With no real data, and for model accuracy while data is collected, third party datasets should be explored. The University of Michigan with Precision Health (2023a;2023b) offers large health datasets with PROMPT Study gathering wearable data. Foschini and Kolbeinsson (2023) used FitBit data collected as part of DiSCover clinical trials.

7.2 Modelling techniques

Rather than SDV, Fan et al. (2020) explored a unified GAN-based framework and found GAN to be promising for relational data synthesis. Given incumbent success, this would be worthwhile exploring for upscaling the dataset for model training.

Neural networks, although more accurate, were not developed due to heavy data requirements, being less interpretable, and computationally expensive. However developing more sophisticated RNNs and LSTMs, known for their ability to process time-series data, is advisable.

Rather than snapshot analysis, the time aspect of the data should be retained and analysed.

7.3 Motion capture smartphone technology

Available mocap technology through smartphone cameras enables analysis of form, joint function and musculoskeletal forces. Uhlich et al. (2023) found OpenCap, developed by Stanford University(2024), to accurately compute skeletal motion and musculoskeletal forces. Similarly, AvaSci(2024) provides means to analyse joint functions. These frontier technologies could increase model reliability, quality, and provide a unique product.

8 Conclusion

This paper has developed and compared a variety of robust ML algorithms recognised by academia for predicting injury risk. The models attempted to predict injury on artificially generated data, personalising them through hyperparameter tuning, with decision tree performing moderately well returning $F2 = 0.581$ and correctly predicting 71.4% injuries. To promote accurate prediction sufficient data quality is required, and artificial creation requires scientific input. Functionality available on Github, provides a good initial framework for Optimi Health given mid-term realisation no real data existed and time remaining.

9 Reference List

1. Ajzen, I., 1991. The theory of planned behavior. *Organizational Behavior and Human Decision Processes* [Online], 50(2). Available from: <https://www.sciencedirect.com/science/article/pii/074959789190020T> [Accessed 24 August 2024].
2. Akilov, O., English, J.C., Ferris, L.K., Ho, J., Moreau, J., Patton, T. and Wolf, J., 2013. Diagnostic Inaccuracy of Smart Phone Applications for Melanoma Detection. *JAMA Dermatology* [Online], 149(4). Available from: <https://jamanetwork.com/journals/jamadermatology/fullarticle/1557488> [Accessed 12 July 2024].
3. Amendolara, A., Bills, K., Donnelly, S., Pfister, D., Settlemayer, M., Shah, M., Wu, V., Johnston, B., Peterson, R., Sant, D. and Kriak, J., 2023. An Overview of Machine Learning Application in Sports Injury Prediction. *Cureus* [Online], 15(9). Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10613321/> [Accessed 16 July 2024].
4. Androutsos, C., Apostolidis, K., Fotiadis, D.I., Mylona, E., Pezoulas, V.C., Tachos, N.S. and Zaridis, D.I., 2024. Synthetic data generation methods in healthcare: A review on open-source tools and methods. *Computational and Structural Biotechnology Journal* [Online], 23. Available from: <https://www.sciencedirect.com/science/article/pii/S2001037024002393> [Accessed 15 August 2024].
5. Arnason, A., Sigurdsson, S.B., Gudmundsson, A., Holme, I., Engebretsen, L. and Bahr, R., 2004. Risk factors for injuries in football. *American Journal of Sports Medicine* [Online], 32. Available from: <https://pubmed.ncbi.nlm.nih.gov/14754854/> [Accessed 15 June 2024].

6. Ayala, F., De Ste Croix, M., Hernandez-Sanchez, S., Lopez-Valenciano, A., Myer, G., Puerta, J.M. and Ruiz-Perez, I., 2018. A Preventive Model for Muscle Injuries : A Novel Approach based on Learning Algorithms. *Medicine & Science in Sports & Exercise* [Online], 50(5). Available from: <https://pubmed.ncbi.nlm.nih.gov/29283933/> [Accessed 16 July 2024].
7. Bai, Y., Garcia, E.A., He, H. and Li, S., 2008. ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning. *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)* [Online], 1-8 June 2008, Hong Kong. Hong Kong: IEEE. Available from: <https://ieeexplore.ieee.org/document/4633969> [Accessed 3 August 2024].
8. Barendrecht, M., Bult, H.J. and Tak, I.J.R., 2018. Injury Risk and Injury Burden Are Related to Age Group and Peak Height Velocity Among Talented Male Youth Soccer Players. *The Orthopaedic Journal of Sports Medicine* [Online], 6(12). Available from: <https://pubmed.ncbi.nlm.nih.gov/30560140/> [Accessed 8 June 2024].
9. Batt, M., Cooper, D. and Palmer, D., 2021. Epidemiology of injury and retirement from sport among former international athletes. *British Journal of Sports Medicine* [Online], 55 (Supplement 1). Available from: https://bjsm.bmj.com/content/55/Suppl_1/A72.3 [Accessed 21 June 2024].
10. Biddle, S.J.H, Chatzisarantis, N.L.D. and Hagger, M.S., 2002. A Meta-Analytic Review of the Theories of Reasoned Action and Planned Behavior in Physical Activity: Predictive Validity and the Contribution of Additional Variables. *Journal of Sport and Exercise Psychology* [Online], 24(1). Available from: <https://journals-humankinetics-com.ezproxy1.bath.ac.uk/view/journals/jsep/24/1/article-p3.xml?content=pdf> [Accessed 23 August 2024].

11. Birişçi, E., Çelik, S. and Gürsakal, N., 2022. *Synthetic Data for Deep Learning: Generate Decision Making and Applications with Python and R* [Online]. Berkely, California: Apress. Available from: <https://learning.oreilly.com/library/view/synthetic-data-for/9781484285879/html/Cover.xhtml> [Accessed 16 August 2024].
12. Blanch, P. and Gabbett, T.J., 2016. Has the athlete trained enough to return to play safely? The acute:chronic workload ratio permits clinicians to quantify a player's risk of subsequent injury. *British Journal of Sports Medicine* [Online], 50(8). Available from: <https://pubmed.ncbi.nlm.nih.gov/26701923/> [Accessed 12 June 2024].
13. Blumkaitis, J.C., During, M., Haller, N., Kranzinger, C., Kranzinger, S., O'Brien, J., Simon, P., Stoggl, T., Strepp, T. and Tomaskovic, A., 2023. Predicting Injury and Illness with Machine Learning in Elite Youth Soccer: A Comprehensive Monitoring Approach over 3 Months. *Journal of Sports Science Medicine* [Online], 22(3). Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10499140/> [Accessed 28 July 2024].
14. Bond, R., Carroll, J.K., Fiscella, K., LeBlanc, W.G., Moorhead, A. and Petrella, R.J., 2017. Who Uses Mobile Phone Health Apps and Does Use Matter? A Secondary Data Analytics Approach. *Journal of Medical Internet Research* [Online], 19(4). Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5415654/> [Accessed 23 August 2024].
15. Bottarelli, M., Cherubin, G., Cohen, S.N., Houssiau, F., Jordon, J., Maple, C., Szpruch, L. and Weller, A., 2022. *Synthetic Data – what, why and how?* [Online]. London: The Royal Society / The Alan Turing Institute. Available from: <https://arxiv.org/abs/2205.03257> [Accessed 16 August 2024].

16. Breiman, L., 2001. Random Forests. *Machine Learning* [Online], 45, pp. 5-32. Available from: <https://link.springer.com/article/10.1023/A:1010933404324> [Accessed 18 July 2024].
17. Califf, R.M., 2018. Biomarker Definitions and their applications. *Experimental Biology and Medicine* [Online], 243. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5813875/> [Accessed 14 July 2024].
18. Caparros, T., Casals, M., Solana, A. and Pena, J., 2018. Low External Workloads Are Related to Higher Injury Risk in Professional Male Basketball Games. *Journal of Sports Science Medicine* [Online], 17(2), pp.289-297. Available from : <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5950746/> [Accessed 25 June 2024].
19. Chandrasekaran, J., Kacker, R.N., Khadka, K., Kuhn, R. and Lei, Y., 2023. Synthetic Data Generation Using Combinatorial Testing and Variational Autoencoder. *2023 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW)* [Online], 16-20 April 2023, Dublin, Ireland. Dublin: IEEE, pp. 228-236. Available from: <https://ieeexplore.ieee.org/document/10132195> [Accessed 17 August 2024].
20. Chen, X. and Yuan, G., 2021. Sports Injury Rehabilitation Intervention Algorithm Based on Visual Analysis Technology. *Mobile Information Systems* [Online], 2021(1). Available from: <https://onlinelibrary.wiley.com/doi/10.1155/2021/9993677> [Accessed 14 September 2024].
21. Chen, Y., Fan, Z., Lan, L., You, L., Zeng, N., Zhang, Z., Zhao, W. and Zhou, X., 2020. Generative Adversarial Networks and Its Applications in Biomedical Informatics. *Frontiers in Public Health* [Online], 8(164). Available from: <https://www.frontiersin.org/journals/public-health/articles/10.3389/fpubh.2020.00164/full> [Accessed 18 August 2024].

22. Cindy, L.Y., Casey, E., Herman, D.C., Katz, N. and Tenforde, A.S., 2018. Sex Differences in Common Sports Injuries. *PM&R* [Online], 10(10), pp. 1073-1082. Available from: <https://onlinelibrary.wiley.com/doi/full/10.1016/j.pmrj.2018.03.008> [Accessed 11 June 2024].
23. Cintia, P., Pappalardo, L. and Rossi, A., 2021. A Narrative Review for a Machine Learning Application in Sports: An Example Based on Injury Forecasting in Soccer. *Sports* [Online], 10(5). Available from: <https://pubmed.ncbi.nlm.nih.gov/35050970/> [Accessed 2 June 2024].
24. Cox, D.R., 1972. Regression Models and Life-Tables. In: Kotz, S., Johnson, N.L. eds. *Breakthroughs in Statistics. Springer Series in Statistics* [Online]. New York: Springer, pp. 527-541. Available from: https://link.springer.com/chapter/10.1007/978-1-4612-4380-9_37 [Accessed 18 July 2024].
25. Coyne, J.O.C., Coutts, A.J., Newton, R.U. and Haff, G.G., 2022. The Current State of Subjective Training Load Monitoring: Follow-Up and Future Directions. *Sports Medicine – Open* [Online], 8(53). Available from: <https://sportsmedicine-open.springeropen.com/articles/10.1186/s40798-022-00433-y> [Accessed 12 September 2024].
26. Datacebo, 2023a. *Synthetic Data Vault: Welcome to the SDV!* [Online]. Available from: <https://docs.sdv.dev/sdv> [Accessed 18 August 2024].
27. Datacebo, 2023b. *Synthetic Data Vault: Synthesizers* [Online]. Available from: <https://docs.sdv.dev/sdv/multi-table-data/modeling/synthesizers> [Accessed 18 August 2024].
28. Datacebo, 2023e. *Synthetic Data Vault: Synthesizers – DayZSynthesizer* [Online]. Available from: <https://docs.sdv.dev/sdv/multi-table-data/modeling/synthesizers/dayzsynthesizer> [Accessed 18 August 2024].

29. Dhanke, J.A., Maurya, R.K., Navaneethan, S., Mavaluru, D., Nuhmani, S., Mishra, N. and Venugopal, E., 2022. Recurrent Neural Model to Analyse the Effect of Physical Training and Treatment in Relation to Sports Injuries. *Computational Intelligence and Neuroscience* [Online], 2022. Available from: [https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9546649/#:~:text=In%20Recurrent%20Neural%20Network%20\(RNN,measure%20to%20avoid%20sports%20injuries.](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9546649/#:~:text=In%20Recurrent%20Neural%20Network%20(RNN,measure%20to%20avoid%20sports%20injuries.) [Accessed 15 September 2024].
30. Domaradzki, J. and Kozlenia, K., 2022. *PeerJ Life & Environment* [Online], 10. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8800385/> [Accessed 12 June 2024].
31. Dovak, J. and Junge, A., 2000. Football injuries and physical symptoms. A review of the literature. *American Journal of Sports Medicine* [Online], 28. Available from: <https://pubmed.ncbi.nlm.nih.gov/11032101/> [Accessed 15 June 2024].
32. Ekstrand, J., Hagglund, M. and Walden, M., 2006a. Previous injury as a risk factor for injury in elite football: a prospective study over two consecutive seasons. *British Journal of Sports Medicine* [Online], 40, pp.767-772. Available from : <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2564391/pdf/767.pdf> [Accessed 15 June 2024].
33. Ekstrand, J., Hägglund, M., Magnusson, H. and Waldén, M., 2016. ACL injuries in men's professional football: a 15-year prospective study on time trends and return-to-play rates reveals only 65% of players still play at the top level 3 years after ACL rupture. *British Journal of Sports Medicine* [Online], 50. Available from: <https://bjsm.bmj.com/content/bjsports/50/12/744.full.pdf> [Accessed 18 June 2024].
34. Ekstrand, J., Walden, M. and Hagglund, M., 2006b. High risk of new knee injury in elite footballers with previous anterior cruciate ligament injury. *British Journal of*

Sports Medicine [Online], 40(2). Available from :
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2492018/#:~:text=Risk%20of%20injury&text=When%20the%20player%20was%20used,%25%20CI%201.8%20to%206.3> [Accessed 15 June 2024].

35. Eliakim, E., Lidor, R., Meckel, Y. and Morgulev, E., 2020. Estimation of injury costs: financial damage of English Premier League teams' underachievement due to injuries. *BMJ Open Sport & Exercise Medicine* [Online], 6. Available from: <https://bmjopensem.bmj.com/content/bmjosem/6/1/e000675.full.pdf> [Accessed 18 June 2024].
36. Fan, J., Chen, J., Du, X., Li, G., Liu, T. and Shen, Y., 2020. Relational Data Synthesis using Generative Adversarial Networks: A Design Space Exploration. *Proceedings of the VLDB Endowment* [Online], 13(12). Available from: <https://www.vldb.org/pvldb/vol13/p1962-fan.pdf> [Accessed 18 August 2024].
37. FDA-NIH Biomarker Working Group 2016. *BEST (Biomarkers, EndpointS, and other Tools) Resource* [Online], Available from: <https://www.ncbi.nlm.nih.gov/books/NBK338448/> [Accessed 14 July 2024].
38. Fisher, K.M., Fuller, L. and Chandler, J.P., 2022. A Review of the Relationship between Heart Rate Monitoring, Training Load, and Injury in Field-Based Team Sport Athletes. *International Journal of Sport Exercise and Health Research* [Online], 6(1), pp.43-54. Available from: https://www.sportscienceresearch.com/IJSEHR_202261_08.pdf [Accessed 3 July 2024].

39. Foschini, L. and Kolbeinsson, A., 2023. Generative models for wearables data. *1st Workshop on Deep Generative Models for Health at NeurIPS 2023* [Online], 15 December 2023. New Orleans, US. NeurIPS. Available from: dd [Accessed 12 September 2024].
40. Gabbe, B.J., Bennell, K.L., Finch, C.F., Wajswelner, H. and Orchard, J.W., 2006. Predictors of hamstring injury at the elite level of Australian football. *Scandinavian Journal of Medicine & Science in Sports* [Online], 16(1). Available from: <https://pubmed.ncbi.nlm.nih.gov/16430675/> [Accessed 23 June 2024].
41. Gabbett, T.J. and Jenkins, D.G., 2011. Relationship between training load and injury in professional rugby league players. *Journal of Science and Medicine in Sport* [Online], 14(3), pp. 204-209. Available from: <https://pubmed.ncbi.nlm.nih.gov/21256078/> [Accessed 13 June 2024].
42. Gabbett, T.J. and Ullah, S., 2012. Relationship between running loads and soft-tissue injury in elite team sport athletes. *Journal of Strength & Conditioning Research* [Online], 26(4), pp.953-960. Available from: <https://pubmed.ncbi.nlm.nih.gov/22323001/> [Accessed 16 June 2024].
43. Gabbett, T.J., 2016. The training-injury prevention paradox: should athletes be training smarter and harder? *British Journal of Sports Medicine* [Online], 50(5). Available from: <https://pubmed.ncbi.nlm.nih.gov/26758673/> [Accessed 12 June 2024].
44. Galarnyk, M., Gouda, P., Quer, G., Steinhubi, S.R. and Topol, E.J., 2020. Inter- and intraindividual variability in daily resting heart rate and its associations with age, sex, sleep, BMI, and time of year: Retrospective, longitudinal cohort study of 92,457 adults. *PLOS One* [Online], 15(2). Available from:

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7001906/> [Accessed 23 August 2024].

45. GeeksforGeeks, 2024. *What are the Advantages and Disadvantages of Random Forest* [Online]. Available from: <https://www.geeksforgeeks.org/what-are-the-advantages-and-disadvantages-of-random-forest/> [Accessed 12 September 2024].
46. Ghazi, K., Wu, S., Zhao, W. and Ji, S., 2021. Instantaneous Whole-Brain Strain Estimation in Dynamic Head Impact. *Journal of Neurotrauma* [Online], 38(8). Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8054523/> [Accessed 14 September 2024].
47. Hagel, B. and Meeuwisse, W.H., 2010. The multi-causality of injury – current concepts. In: van Mechelen, W. and Verhagen, E. *Sports Injury Research*. New York: Oxford University Press, ch.8.
48. Hajratalli, N.K., Henson, P., Onnela, J., Torous, J., Vaidyam, A. and Wisniewski, H., 2019. Understanding the quality, effectiveness and attributes of top-rated smartphone health apps. *British Medical Journal Mental Health* [Online], 22 (1). Available from: <https://mentalhealth.bmj.com/content/ebmental/22/1/4.full.pdf> [Accessed 13 July 2024].
49. <https://www.sciencedirect.com/science/article/pii/S0166497222001456#bib133>
50. Huang, K. and Ihm, J., 2021. Sleep and Injury Risk. *Current Sports Medicine Reports* [Online], 20(6), pp.286-290. Available from: <https://pubmed.ncbi.nlm.nih.gov/34099605/> [Accessed 2 July 2024].
51. Huang, S. Cai, N., Pacheco, P.P., Narandes, S., Wang, Y. and Xu, W., 2018. Applications of Support Vector Machine (SVM) Learning in Cancer Genomics. *Cancer Genomics & Proteomics* [Online], 15(1). Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5822181/> [Accessed 27 July 2024].

52. Imbalanced-learn developers, 2024. *ADASYN – Version 0.12.3* [Online]. Imbalanced-learn. Available from: https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.ADASYN.html [Accessed 23 August 2024].
53. Jayanthi, N., Kleithernes, S., Dugas, L., Pasulka, J., Iqbal, S. and LaBella, C., 2020. Risk of Injuries Associated With Sport Specialization and Intense Training Patterns in Young Athletes: A Longitudinal Clinical Case-Control Study. *The Orthopaedic Journal of Sports Medicine* [Online], 8(6). Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7318830/> [Accessed 19 June 2024].
54. Jiang, L.C. and Mengru, S., 2022. Linking social features of fitness apps with physical activity among Chinese users: Evidence from self-reported and self-tracked behavioral data. *Information Processing & Management* [Online], 59(6). Available from: <https://www.sciencedirect.com/science/article/pii/S0306457322001972> [Accessed 24 August 2024].
55. Jones, B.H., Hauret, K.G., Dye, S.K., Hauschild, V.D., Rossi, S.P., Richardson, M.D. and Friedl, K.E., 2017. *Journal of Science and Medicine in Sport* [Online], 20(4). Available from : <https://www.sciencedirect.com/science/article/pii/S1440244017310617#bibl0005> [Accessed 12 June 2024].
56. Kallinen, M. and Markku, A., 1995. Aging, Physical Activity and Sports Injuries: An Overview of Common Sports Injuries in the Elderly. *Sports Medicine* [Online], 20(1), pp.41-52. Available from: <https://link.springer.com/article/10.2165/00007256-199520010-00004> [Accessed 8 June 2024].
57. Kautz, T., Groh, B.H., Hannink, J., Jensen, U., Strubberg, H. and Eskofier B.M., 2017. Activity Recognition in beach volleyball using a Deep Convolutional Neural Network.

- Data Mining and Knowledge Discovery* [Online], 31, pp.1678-1705. Available from: <https://link.springer.com/article/10.1007/s10618-017-0495-0> [Accessed 12 September 2024].
58. Knapik, J.J., Swedler, D.I., Grier, T.L., Hauret, K.G., Bullock, S.H., Williams, K.W., Darakjy, S.S., Lester, M.E., Tobler, S.K. and Jones, B.H., 2009. Injury reduction effectiveness of selecting running shoes based on plantar shape. *Journal of Strength and Conditioning Research* [Online], 23(3), pp.685-697. Available from: <https://www.scopus.com/record/display.uri?eid=2-s2.0-68049137517&origin=inward&txGid=05e9ffd91d0d87a17baa383298a46177> [Accessed 20 June 2024].
59. Kucera, K.L., Marshall, S.W., Kirkendall, D.T., Marchak, P.M. and Garrett Jr., W.E., 2005. Injury history as a risk factor for incident injury in youth soccer. *British Journal of Sports Medicine* [Online], 39(7). Available from: <https://pubmed.ncbi.nlm.nih.gov/15976172/> [Accessed 15 June 2024]
60. Kumar, G.S., Kumar, M.D., Reddy, S.V.R., Kumari, B.V.S. and Reddy, C.R., 2024. Injury Prediction in Sports using Artificial Intelligence Applications: A Brief Review. *Journal of Robotics and Control (JRC)* [Online], 5(1). Available from: <https://journal.umy.ac.id/index.php/jrc/article/view/20814> [Accessed 25 July 2024].
61. Lobery, L.A., 2009. *In That Instant It Was Over: The Athlete's Experience of a Career-Ending Injury* [Online]. Dissertation (PhD). University of Tennessee. Available from: https://trace.tennessee.edu/cgi/viewcontent.cgi?referer=&httpsredir=1&article=1085&context=utk_graddiss [Accessed 22 June 2024].
62. Makhni, E.C., Lee, R.W., Nwosu, E.O., Steinhaus, M.E. and Ahmad, C.S., 2015. Return to competition, re-injury, and impact on performance of preseason shoulder injuries in Major League Baseball pitchers. *The Physician and Sportsmedicine*

[Online], 43(3). Available from:
<https://www.tandfonline.com/doi/full/10.1080/00913847.2015.1050952#abstract>

[Accessed 24 June 2024].

63. Malkov, D. and Kim, J., 2022. The Application of Machine Learning on the Injury Prediction of Soccer Players. *CEUR Workshop Proceedings: 1st International Workshop on Intelligent Software Engineering, 2022* [Online]. Available from: https://ceur-ws.org/Vol-3362/ISE2022_short02_Malikov_Application.pdf [Accessed 25 June 2024].
64. Matzkin, E. and Garvey, K., 2019. Sex Differences in Common Sports-Related Injuries. *NASN School Nurse* [Online], 34(5), pp.266-269. Available from: <https://journals.sagepub.com/doi/abs/10.1177/1942602X19840809> [Accessed 10 June 2024].
65. Maupin, D., Schram, B., Canetti, E. and Orr, R., 2020. The Relationship Between Acute: Chronic Workload Ratios and Injury Risk in Sports: A Systematic Review. *Open Access Journal of Sports Medicine* [Online], 11, pp.51-75. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7047972/> [Accessed 28 June 2024].
66. Meeuwisse, W.H., 1994. Athletic Injury Etiology: Distinguishing Between Interaction and Confounding. *Clinical Journal of Sport Medicine* [Online], 4(3). Available from: https://journals.lww.com/cjsportsmed/abstract/1994/07000/athletic_injury_etiology_distinguishing_between.5.aspx [Accessed 14 July 2024].
67. Meng, L. and Qiao, E., 2021. Analysis and design of dual-feature fusion neural network for sports injury estimation model. *Neural Computing and Applications* [Online], 35, pp. 14627-14639. Available from: <https://link.springer.com/article/10.1007/s00521-021-06151-y> [Accessed 14 September 2024].

68. Naidu, G., Zuva, T. and Sibanda, E.M., 2023. A Review of Evaluation Metrics in Machine Learning Algorithms. In: Silhavy, R., Silhavy, P. (eds) *Artificial Intelligence Application in Networks and Systems* [Online]. CSOC 2023. pp. 15-25. Available from: https://link.springer.com/chapter/10.1007/978-3-031-35314-7_2 [Accessed 5 September 2024].
69. National Center for Health Statistics (NCHS), 2018a. *National Health Interview Survey (NHIS) 2017 Data Release: Sample Adult file, CSV data* [Online]. National Center for Health Statistics. Available from: https://www.cdc.gov/nchs/nhis/nhis_2017_data_release.htm [Accessed 23 August 2024].
70. National Center for Health Statistics (NCHS), 2018b. *National Health and Nutrition Examination Survey (NHANES) 2017-2018* [Online]. National Center for Health Statistics. Available from: <https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?BeginYear=2017> [Accessed 14 August 2024].
71. National Center for Health Statistics (NCHS), 2020a. *National Health and Nutrition Examination Survey (NHANES) 2017-2018 Data Documentation, Codebook, and Frequencies – Demographic Variables and Sample Weights (DEMO_J)* [Online]. National Center for Health Statistics. Available from: https://wwwn.cdc.gov/Nchs/Nhanes/2017-2018/DEMO_J.htm [Accessed 14 August 2024].
72. National Center for Health Statistics (NCHS), 2020b. *National Health and Nutrition Examination Survey (NHANES) 2017-2018 Data Documentation, Codebook, and Frequencies – Body Measures (BMX_J)* [Online]. National Center for Health

Statistics. Available from: https://wwwn.cdc.gov/Nchs/Nhanes/2017-2018/BMX_J.htm [Accessed 14 August 2024].

73. Neto, F.R., Tibana, R.A., Dorneles, R. and Costa, R.R.G., 2022. Internal and External Training Workload Quantification in 4 Experienced Paracanoeing Athletes. *Journal of Sports Rehabilitation* [Online], 31(2), pp.239-245. Available from: <https://pubmed.ncbi.nlm.nih.gov/34426553/> [Accessed 26 June 2024].
74. NICE, 2024. *Evidence Standards Framework for Digital Health Technologies: User Guide* [Online]. Available from <https://www.nice.org.uk/Media/Default/About/what-we-do/our-programmes/evidence-standards-framework/user-guide.pdf> [Accessed 5 July 2024].
75. Pappalardo, L., Guerrini, L., Rossi, A. and Cintia, P., 2019. Explainable Injury Forecasting in Soccer via Multivariate Time Series and Convolutional Neural Networks. *Barca Innovation Hub: Barca Sports Analytics Summit 2019* [Online]. Camp Nou, Barcelona: FC Barcelona, 13 November 2019. Available from: https://static.capabiliaserver.com/frontend/clients/barca/wp_prod/wp-content/uploads/2020/01/c6658839-paper-format-luca-pappalardo-1.pdf [Accessed 11 September 2024].
76. Pasqualini, I., Ranalletta, M., Rossi, L.A. and Tanoira, I., 2022. Factors That Influence the Return to Sport After Arthroscopic Bankart Repair for Glenohumeral Instability. *Open Access Journal of Sports Medicine* [Online], 13. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8985826/> [Accessed 18 June 2024].
77. Patki, N., Veeramachaneni, K. and Wedge, R., 2016. The Synthetic Data Vault. *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)* [Online], 17-19 October 2016, Montreal, Canada. Montreal: IEEE. Available from: <https://ieeexplore.ieee.org/document/7796926> [Accessed 12 August 2024].

78. Pawar, B., 2024. *Fortune Business Insights Industry Reports: mHealth market* [Online]. US: Fortune Business Insights. Available from: <https://www.fortunebusinessinsights.com/industry-reports/mhealth-market-100266> [Accessed 2 July 2024].
79. Pons-Villanueva, J., Segui-Gomez, M. and Martinez-Gonzalez, M.A., 2010. Risk of injury according to participation in specific physical activities: a 6-year follow-up of 14 356 participants of the SUN cohort. *International Journal of Epidemiology* [Online], 39(2), pp.580-587. Available from: <https://academic.oup.com/ije/article/39/2/580/679411> [Accessed 18 June 2024].
80. Porta, M., 2016. *A Dictionary of Epidemiology* [Online], 6th ed. Oxford: Oxford University Press. Available from: <https://www-oxfordreference-com.ezproxy1.bath.ac.uk/display/10.1093/acref/9780199976720.001.0001/acref-9780199976720> [Accessed 16 July 2024].
81. Reynolds, K.L., Heckel, H.A., Witt, C.E., Martin, J.W., Pollard, J.A., Knapik, J.J. and Jones, B.H., 1994. Cigarette Smoking, Physical Fitness, and Injuries in Infantry Soldiers. *American Journal of Preventive Medicine* [Online], 10(3), pp.145-150. Available from: <https://www.sciencedirect.com/science/article/abs/pii/S074937971830610X> [Accessed 21 June 2024].
82. Ruddy, J.D., Shield, A.J., Maniar, N., Willaims, M.D., Duhig, S., Timmins, R.G., Hickey, J., Bourne, M.N. and Opar, D.A., 2018. Predictive Modeling of Hamstring Strain Injuries in Elite Australian Footballers. *Medicines & Sciences in Sports & Exercise* [Online]. 50(5). Available from: https://journals.lww.com/acsm-msse/fulltext/2018/05000/predictive_modeling_of_hamstring_strain_injuries.4.aspx [Accessed 28 July 2024].

83. Sayad, S., 2024 *OneR*. Available from: <https://www.saedsayad.com/oner.htm> [Accessed 12 July 2024].
84. Schmitz, M., 2022. *Health apps: these legal hurdles must be observed* [Online]. Oppenhoff. Available from: <https://www.oppenhoff.eu/en/news/detail/health-apps-these-legal-hurdles-must-be-observed/> [Accessed 3 June 2024].
85. Scikit-learn developers, 2024a. *Scikit-learn: 3. Model selection and evaluation – 3.2. Tuning the hyper-parameters of an estimator* [Online]. Scikit-learn. Available from: https://scikit-learn.org/stable/modules/grid_search.html#grid-search [Accessed 16 September 2024].
86. Scikit-learn developers, 2024b. *Scikit-learn: Support Vector Machines – Plot classification boundaries with different SVM kernels* [Online]. Scikit-learn. Available from: https://scikit-learn.org/dev/auto_examples/svm/plot_svm_kernels.html [Accessed 17 September 2024].
87. Scikit-learn developers, 2024c. *Scikit-learn: 1. Supervised learning – Decision Trees* [Online]. Available from: <https://scikit-learn.org/stable/modules/tree.html#tree> [Accessed 20 July 2024].
88. Scikit-learn developers, 2024d. *Scikit-learn: 3. Model Selection and Evaluation – Cross-validation: evaluating estimator performance* [Online]. Available from: https://scikit-learn.org/stable/modules/cross_validation.html#stratified-k-fold [Accessed 12 September 2024].
89. Scikit-learn developers, 2024e. *Scikit-learn: 7. Dataset loading utilities – Generated datasets* [Online]. Available from: https://scikit-learn.org/dev/datasets/sample_generators.htm [Accessed 12 September 2024].
90. Scikit-learn developers, 2024f. *Scikit-learn: HalvingGridSearchCV* [Online]. Available from: https://scikit-learn.org/dev/api/modules/generated/sklearn.model_selection.HalvingGridSearchCV.html

learn.org/stable/modules/generated/sklearn.model_selection.HalvingGridSearchCV.html#sklearn.model_selection.HalvingGridSearchCV [Accessed 12 September 2024].

91. Senbel, S., Sharma, S., Raval, M.S., Taber, C., Nolan, J., Artan, N.S., Ezzeddine, D. and Kaya, T., 2021. Impact of Sleep and Training on Game Performance and Injury in Division-1 Women's Basketball Amidst the Pandemic. *IEEE Access* [Online], 10, pp. 15516-15527. Available from: <https://ieeexplore.ieee.org/document/9690164> [Accessed 2 September 2024].
92. Singh, J., Singh, G. and Singh, R., 2017. Optimization of sentiment analysis using machine learning classifiers. *Human-centric Computing and Information Sciences* [Online], 7(32). Available from: <https://hcis-journal.springeropen.com/articles/10.1186/s13673-017-0116-3> [Accessed 15 July 2024].
93. Smartdraw, 2024. *Decision Tree* [Online]. Available from: <https://www.smartdraw.com/decision-tree/> [Accessed 24 September 2024].
94. Stanford University, 2024. *OpenCap: Musculoskeletal forces from smartphone video* [Online]. Available from: <https://www.opencap.ai/> [Accessed 12 September 2024].
95. Statista, 2023. *Statistics report on the global m-health industry and market* [Online]. Available from: <https://www.statista.com/study/24501/mhealth-statista-dossier/> [Accessed 3 July 2024].
96. Uhlrich, S.D., Falisse, A., Kidzinski, L., Muccini, J., Ko, M., Chaudhari, A.S., Hicks, J.L. and Delp, S.L., 2023. OpenCap: Human movement dynamics from smartphone videos. *PLoS Computational Biology* [Online], 19(10). Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10586693/> [Accessed 2 September 2024].

97. UK Government, 2017. *Criteria for health app assessment* [Online]. Available from: <https://www.gov.uk/government/publications/health-app-assessment-criteria/criteria-for-health-app-assessment#:~:text=All%20apps%20must%20work%20and,clinical%20experience%20and%20patient%20preferences>. [Accessed 4 July 2024].
98. Ullah, S., Gabbett, T.J. and Finch, C.F., 2012. Statistical modelling for recurrent events: an application to sports injuries. *British Journal of Sports Medicine* [Online], 48(17). Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4145455/> [Accessed 25 July 2024].
99. University of Michigan (UMich), 2023a. *Precision Health Documentation – PROMPT* [Online]. Available from: <https://sites.lsa.umich.edu/phanalyticsdocumentation/prompt/> [Accessed 17 September 2024].
100. University of Michigan (UMich), 2023b. *Precision Health – Data Access & Tools, Datasets* [Online]. Available from: <https://precisionhealth.umich.edu/data-access-tools/datasets/> [Accessed 17 September 2024].
101. Vafeiadis, T., Diamantaras, K.I., Sarigiannidis, G. and Chatzisavvas, K.Ch., 2015. A comparison of machine learning techniques for customer churn prediction. *Simulation Modelling Practice and Theory* [Online], 55. Available from: <https://www.sciencedirect.com/science/article/pii/S1569190X15000386?via%3Dihub#s0040> [Accessed 20 September 2024].
102. Van Eetvelde, H., Ley, C., Mendonca, L. D., Seil, R. and Tischer, T., 2021. Machine Learning Methods in Sport Injury Prediction and Prevention: A Systematic Review. *Journal of Experimental Orthopaedics* [Online], 8(27). Available from:

<https://jeo-esska.springeropen.com/articles/10.1186/s40634-021-00346-x>

[Accessed 25 July 2024].

103. Viegas, F., Ocarino, J.M., Freitas, L.S., Pinto, M.C., Facundo, L.C., Amaral, A.S., Silva, S., de Mello, M.T. and Silva, A., 2022. The sleep as a predictor of musculoskeletal injuries in adolescent athletes. *Sleep Science* [Online], 15(3), pp. 305-311. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9496483/#:~:text=Sleep%20is%20essential%20for%20musculoskeletal,predispose%20them%20to%20sports%20injuries>. [Accessed 2 July 2024].
104. Von Jouanne-Diedrich, H.K., 2017. OneR - Establishing a New Baseline for Machine Learning Classification Models [Online]. CRAN. Available from: <https://cran.r-project.org/web/packages/OneR/vignettes/OneR.html> [Accessed 12 July 2024].
105. Zech, A., Hollander, K., Junge, A., Steib, S., Groll, A., Heiner, J., Nowak, F., Pfeiffer, D. and Rahlf, A.L., 2021. Sex differences in injury rates in team-sport athletes: A systematic review and meta-regression analysis. *Journal of Sport and Health Science* [Online], 11(1), pp. 104-114. Available from: <https://www.sciencedirect.com/science/article/pii/S2095254621000545> [Accessed 10 June 2024].
106. Zoellner, A., Whatman, C., Sheerin, K. and Read, P., 2022. Prevalence of sport specialisation and association with injury history in youth football. *Physical Therapy in Sport* [Online], 58, pp.160-166. Available from: <https://www.sciencedirect.com/science/article/pii/S1466853X22001456> [Accessed 18 June 2024].

107. Zumeta-Olaskoaga, L., Weigert, M., Larruskain, J, Bikandi, E., Setuain, I., Lekue, J., Kuchenhoff, H. and Lee, D., 2021. Prediction of sports injuries in football: a recurrent time-to-event approach using regularized Cox models. *AStA Advances in Statistical Analysis* [Online], 107. Available from: <https://link.springer.com/article/10.1007/s10182-021-00428-2> [Accessed 25 July 2024].

10 Appendix

10.1 Biomarker Literature Review

The aim of this avenue of the literature review is to identify biomarkers highlighted by the academic community in their role as determinants of injury risk for future development of the Optimi Health app and model.

Biomarkers are defined as characteristics “*measured as an indicator of normal biological processes, pathogenic processes, or biological responses to an exposure or intervention, including therapeutic interventions...*” (FDA-NIH Biomarker Working Group, 2016). Given the nature of a mobile application, much data appropriate to this definition is not recordable by users, i.e. thermal imaging, functional screen movement or blood samples to name a few. Instead, the model inputs have to account for ease of user data recording, limiting the scope of the model inputs possible.

Hagel and Meeuwisse (2010) highlighted the need to comprehend the sports-injury aetiology to identify relevant biomarkers. Meeuwisse’s (1994) multi-factorial model of injury acknowledges the complexity of injury risk prediction. A key facet is the manner in which various factors combine to contribute to cause injury, whether they be **interaction** or **confounding**. Confounding facets vary among studies and hence requires control, yet interaction ones are a function of the underlying nature of injury (Meeuwisse, 1994).

Meeuwisse’s model provides a basis for splitting injury into three broad areas: intrinsic factors, extrinsic factors and an inciting event. Intrinsic factors involve personal non-modifiable aspects inherent to an individual. They involve personal statistics such as age, physical health such as injury history, also anatomy such as joint alignment, arch

and biomechanical factors, i.e. flexibility or muscle strength. Extrinsic factors entail non-physical elements of an individual, including their training routine and the training loads exerted, the sports they undertake and demands placed on body parts through this to the weather conditions and equipment used i.e. quality of shoe for arch.

Combinations of these intrinsic and extrinsic factors can bring about a more injury susceptible athlete, representing increased likelihood of an inciting event. The inciting event represents the final necessary cause such as certain force and moments exerted on a joint at one time i.e. pivot motion leading to ACL tear. Bahr and Holme (2003) highlight that it is the interaction and confounding nature of these risks that culminate in the injury prone outcome.

Cintia et al. (2021) summarise that personal statistics such as age, gender, weight, height and resultant body mass index (BMI) are common characteristics used to develop injury prediction models. Domaradzki and Kozlenia (2022) highlight the significance of different body composition metrics while multiple studies on youth injuries note the effect of age, weight, and height.

Barendrecht et al. (2018) identified younger athletes to face elevated injury risk due to their bodies still developing. Their findings align with findings attaining to growth reducing flexibility and bone density while size effects on coordination pose additional risk, whereby injury 6 months after peak height velocity was 31% above overall mean. Meanwhile, Kallinen and Markku (1995) summarise that aging into the elderly years poses an increased risk of injury, in both acute and overuse injuries. The degenerative aging processes causing structural and functional body changes with connective tissue stiffening, putting muscles and tendon tissues at greater risk while loss of minerals in bones reduces their mechanical strength. Meanwhile, chronic overuse from long term cumulative effects of physical activity exacerbated by aging is more

common. Although, they recognise sufficient levels of physical activity in aging individuals can ensure maintenance of physical and functional capabilities mitigating injury risk. These aspects signify how age and physical activity can act as interacting factors with physical activity modulating the risk posed by age.

Gender is important to monitor with the effects of sports on the injury incidence found to vary between sexes. Matzkin and Garvey (2019) provide in-depth justifications starting with context of the National Institutes of Health (NIH) implementing changes to ensure increased participation of females in their clinical trials after recognising the significance of sex differences in health. Soon after, investigations found sex-based differences within the musculoskeletal system such as males possessing greater bone density, muscle mass, and lean mass. Matzkin and Garvey (2019) further explain the extent to which and why females are generally more susceptible to ACL injuries. As well, Zech et al. (2021) highlights that there appear to be sex-specific differences with male team sport players experiencing more overall, upper extremity, hip/groin, thigh and foot injuries than female counterparts while female players experienced more ACL injuries. They acknowledge that sex should be considered in injury risk analysis in case of an interaction effect with other variables, i.e. main sport undertaken, as further exaggerated by Cindy et al. (2018) in their comparison of the sexes and injury and by Domaradzki and Kozlenia (2022) in finding different body composition metrics more appropriate per sex for injury risk determination.

The body composition of individuals is found to be influential intrinsic factor in the aetiology of injury risk. Domaradzki and Kozlenia (2022) define cut off points for body composition metrics, finding men with body mass index (BMI) over 24.38 and skeletal mass index (SMI) over 16.40, and women with muscle to fat ratio (MFR) over 1.67 and fat mass index (FMI) over 4.17 are more likely to be injured. Further identifying

biological differences between the sexes as well as pointing out that mass is not necessarily a sufficient measure, but fat, muscle and skeletal mass are significant distinctions in overall mass. Meanwhile, in their study of army trainees, Jones et al. (2017) identified a higher risk of injury to trainees with the lowest BMIs and lowest levels of aerobic fitness (two-mile run times), while lowest injury risk found among most aerobically fit who exhibited “average” weight or over-average levels. This highlights the interaction effect of aerobic fitness combined with a necessary balance of mass given the individual’s height and a more general idea that greater physical fitness and healthier bodies are important in mitigating injury risk. Although these two provide slightly mixed stories, it is clear height and body mass (and its granular component metrics) can be related to injury risk.

History of prior injuries has been identified as one of the primary risk factors individuals face to injury. Multiple studies have identified prior injury as being linked to a successive injury of the same body part, as summarised in Butler et al. (2014) systematic review. Ekstrand et al.’s (2006a) study was consistent with prior studies in identifying previous injury as a significant risk factor for injury in football players, with a positive relation between the number of prior injuries and re-injury (Arnason et al., 2004; Dvorak et al., 2000; Kucera et al., 2005). They previously had found that resultant deficits in physical condition, or altered functional movement after a prior injury may provide a causal link to a successive anatomically unrelated injury, i.e. a prior ACL injury was found to increase risk of a new knee injury (Ekstrand et al., 2006b). This means that further investigation into how an injury in one body part could potentially increase forces and moments placed on other body parts may be useful – possibly investigative via motion capture technology. A multitude of other studies have

found similar results as the ACL case, such as in hamstring and shoulder injuries (Gabbe et al., 2006; Makhni et al., 2015)

Cintia et al. (2021) used two main categories of data input to predict injuries: Training workloads and players' psychophysiological assessment.

Training workloads is further split into external and internal workloads. External workloads entail training features that describe the effort performed during physical activity. Meanwhile, internal workload involves the perceived exertion of the individual from said physical activity session (Neto et al., 2022).

Despite common perceptions that high external workloads (i.e. movements, accelerations) increases the risk of injury, evidence exists that such high training can resemble preventative efforts against injury. Caparros et al. (2018), in their study of basketballers, found reduced kinematic workload – fewer decelerations and less distance covered – were significantly associated with injury during games. This reinforces previous findings that a minimum chronic external workload is important in preventing injuries (Blanch and Gabbett, 2016; Gabbett and Jenkins, 2011)

Gabbett (2016) further supports this yet highlights that considering all 'high training loads' as carrying equal injury risk is inappropriate due to the variance in which they can be attained. Differences occur via volume, intensity, frequency and balance of training activities etc. He singles out findings from (Gabbett and Ullah, 2012) whereby greater amounts of very-high speed running were associated with higher injury risk, yet with players performing greater amounts of low-intensity activity and short acceleration efforts there risk was lower. This is suggestive that appropriate high training load programmes can be a source of preventative means. Cintia et al. (2021) state that many metrics for external workload can be recorded via GPS, providing

access to kinematic, metabolic and mechanical data. Possibly available via wearable technology, GPS could track an individual's overall movement during a training session and at what speeds, energy expenditure in the session and muscular-skeletal load i.e. explosive distance, accelerations and decelerations. Hence, training workloads, particularly their aggregation into chronic and acute workloads provides an instrumental point of view, as commonly noted for the ACWR metric (Maupin et al., 2020)

The sport individuals undertake and the degree of specialism in sports can be considered as significant in affecting injury risk. Different sports encourage stress on different parts of the body, such as tennis often being associated with knee, shoulder and elbow injuries while football faces common injuries in the upper leg, knee, lower leg, ankle and feet. Pons-Villanueva et al. (2009) use the Cox proportional hazards regression on a long run study of a large sample to evaluate injury risk associated with participation in various sports. They identified football, other team sports and athletics to demonstrate strong harmful associations for injuries among men while among women team sports and skiing were associated with the highest risk among women. Concurrently, Zoellner et al. (2022) and Jayanthi et al. (2020) identified that the degree of sport specialisation significantly increased injury risk among young athletes.

Other notable metrics that might be useful for injury prediction include:

- Equipment (Knapik et al., 2009)
- Smoking (Reynolds et al., 1994)
- Psychophysiological assessment (Ayala et al., 2018)
- Sleep (Viegas et al., 2022; Huang and Ihm, 2021)
- Resting heart rate (Fisher et al., 2022)

- Motion capture technology to capture joint and musculoskeletal functionality
(Uhlrich et al., 2023)

10.2 Synthetic Data Vault (SDV) Methodology

Given interacting variables, the construction of an SDG that could handle relational data was important. The majority look at single table generation rather than multi-table, limiting the use of popular GANs and VAEs at this time (Chen et al., 2020).

The Synthetic Data Vault (SDV) library provides a full means through its “recursive conditional parameter aggregation” (Patki, Veeramachaneni and Wedge, 2016). The conditional parameter aggregation (CPA) method is sufficient due to only two generations here. It specifies how child tables must be incorporated into the parent. The free-user HMASynthesiser requires two inputs: the original dataset and the metadata – establishing table variables and the relations between tables, via primary and foreign keys.



Figure 10.2a: Simplified CPA process (Patki et al.,2016).

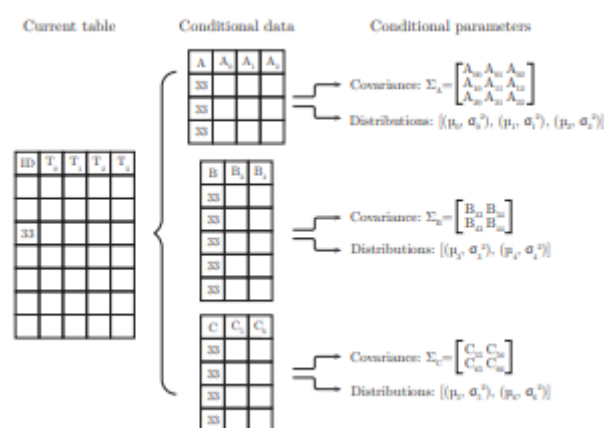



Figure 10.2b: Example of CPA for a row iteration with primary key “33” (Patki et al.,2016).

The CPA is composed of four steps:


- 1) First, iterates through each row of the parent table (primary)
- 2) Generates sets of conditional data for each child table containing the primary key (user id) – m different foreign key columns referring to current table $\rightarrow m$ sets of conditional data
- 3) For each set of conditional data, performs multivariate Gaussian Copula process. This describes the joint distribution of multiple random variables, yielding m sets of distributions and m sets of covariance matrices, known as *conditional parameters*, as Figure 10.2b
- 4) These are inputted as new “derived columns” in the parent table, as each row value represents subset of data from a child table given the user id, generating an extended table – as Figure 10.2c

Extended table


ID	T ₀	T ₁	T ₂	T ₃	A ₀₀	A ₀₁	...	C ₀₀	H ₀	σ_0^2	...	σ_6^2
33												



Original columns



Covariances



Distributions

Figure 10.2c: Result of CPA (Patki et al.,2016).

The use of RCPA is not necessary for this project as child tables do not have their own children, with all tables being related to a single primary parent table. The programme then attains the statistical properties of each variable, and between variables, calculating *cdf* functions and covariances through applying the Gaussian Copula on the extended table.

The *cdf* and covariances form the generative model and synthetic data is generated by sampling data from these calculated distributions and covariances. Once sampled, can factor in primary and foreign key relations to synthesise the tables, and the database.¹

¹ Note the specifics of how different data types are processed to make the model and generated to attain the synthetic dataset can be found in greater detail in the original paper.

10.3 OneR Model Method

The OneR model is trained via this brief process (Sayad,2024)

- 1) For each covariate,
 - For each value of covariate:
 - Count frequency of each target value
 - Find most frequent target class
 - Make the rule that assigns that class to this covariate value
 - Calculate the total error of the rules of each covariate
- 2) Choose the predictor with the smallest total error

10.4 Model Evaluative Metrics

For each trained and tested model, a host of metrics are calculated to provide a broader overall picture of their performance. The aim of the models is to predict injury imminent cases. Hence, there is special focus trained on correctly predicting injury imminent cases, prioritised over predicting non-injury cases – hence appropriate cost matrix applied. As in section 4.7.1, special attention is applied to Recall and F2 scores. The metrics assessed, but not defined in section 4.7.1 are:

- Specificity
- Precision and Negative Predicted Value (NPV)
- F1-score
- AUC-ROC

There are four key classification counts devised from the confusion matrix:

- True positive (TP) represents positive instances correctly predicted as positive
- False positive (FP) represents negative instances incorrectly predicted as positive
- True negative (TN) represents negative instances correctly predicted as negative
- False negative (FN) represents positive instances incorrectly predicted as negative

Naidu et al. (2023) provide an insightful review of evaluation techniques. I use their analysis and that of Cintia et al. (2021) to aid my approach.

Specificity

Specificity resembles the same metric as Recall (sensitivity) but for negative instances. It represents the model's ability to predict a true negative instance correctly:

$$Specificity = \frac{TN}{TN + FP}$$

Precision

Precision measures the ratio of correct predictions. It calculates the ratio of how many predicted positive instances are actual positive instances. This can be useful in indicating if the model is prioritising classification of positive instances.

$$Precision = \frac{TP}{TP + FP}$$

$$NPV = \frac{TN}{TN + FN}$$

A low score is fine if the cost of misclassifying a negative instance is low. A high score alone is not sufficient, as the metric does not account for how many actual positive instances were predicted negative. Hence, the use of the recall score which focuses on the degree to which positive instances are correctly classified. A reasonable precision score is useful, however, as Optimi Health do not want to be creating too many unnecessary prevention plans. NPV is the same as precision, but for negative instances.

F1-score

As mentioned above, a combination of precision and recall scores are useful in evaluating the performance of the model. The F1-score metric is the harmonic mean of the precision and recall scores, applying a balanced weighting to both scores to provide an overall metric (Vafeiadis et al. 2015).

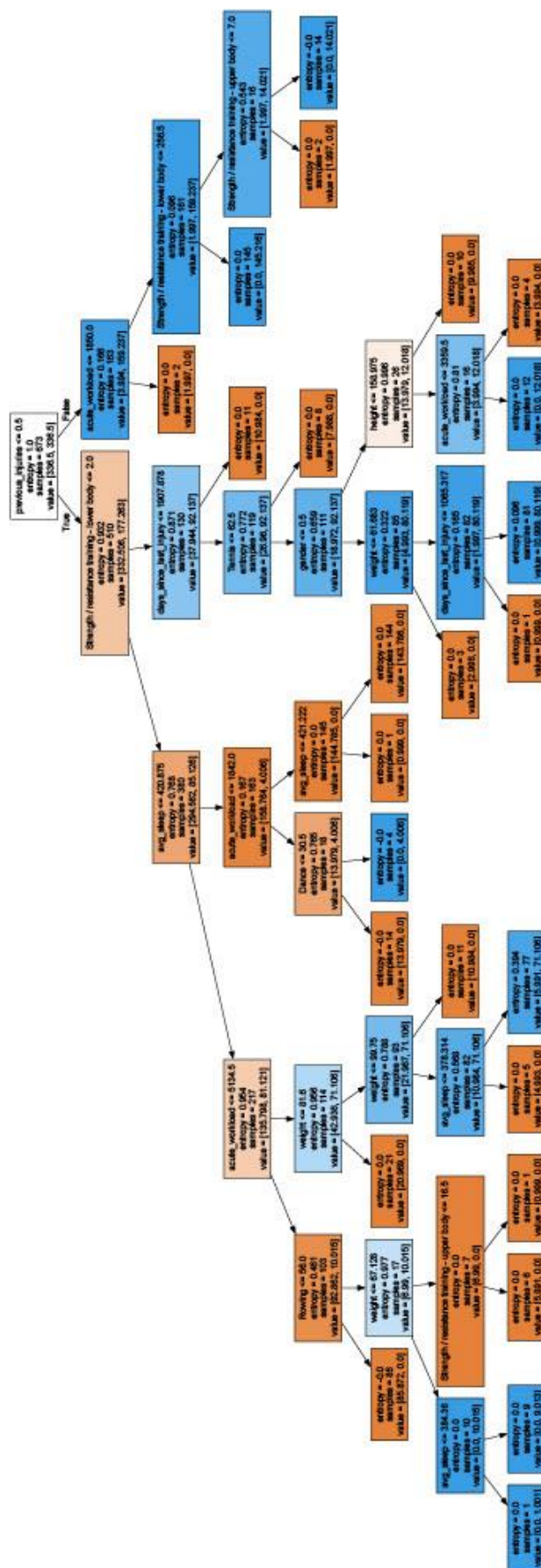
$$F1\ Score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

However, this assumes precision and recall are equally important. As above, recall score is more important. Although false positives may create extra work for Optimi Health in terms of preventative plan generation, false negatives means that Optimi is not predicting injury instances as well. Hence, there is a greater importance placed on reducing false negatives over false positives – and the recall score gives better indication of this ($Recall = \frac{TP}{TP+FN}$).

AUC-ROC

Area under the Curve (AUC) metric provides an aggregate performance measure. It represents the probability that a random positive instance is ranked higher than a randomly negative one by the model by plotting the true positive rate against the false positive rate. A higher AUC value suggests better model performance.

10.5 Decision Tree Model Graphic



10.6 Models with hyperparameter tuning using Recall scores

The models used a 70:30 training-test split for the 650-user dataset. Following ADASYN on training data only, to avoid overfitting, models were trained on 348 injury and 480 non-injury instances using grid search techniques, tuning hyperparameters according to recall score – prioritising injury prediction while acknowledging non-injuries. Five-fold SCV evaluates the predictive performance of the different algorithms in Figure 10.6.1.

Figure 10.6.1: Full evaluative metrics for models, tuned on recall scores — five-fold stratified cross-validation (ADASYN)								
Model	Class	Recall	Precision	F1-score	F2-score	AUC-ROC	Predicted	Misspecified
Baseline	No injury imminent	0.469	0.482	0.475	0.471	0.485	158	179
	Injury imminent	0.501	0.489	0.495	0.499	0.485	171	170
Custom OneR	No injury imminent	0.994	0.788	0.879	0.945	0.844	335	2
	Injury imminent	0.736	0.992	0.845	0.776	0.844	251	90
Decision Tree	No injury imminent	0.958	0.961	0.960	0.959	0.969	323	14
	Injury imminent	0.962	0.959	0.960	0.961	0.969	328	13
Random Forest	No injury imminent	0.994	0.957	0.975	0.986	0.999	335	2
	Injury imminent	0.956	0.994	0.975	0.963	0.999	326	15
Support Vector Machine	No injury imminent	0.404	0.500	0.447	0.420	0.497	136	201
	Injury imminent	0.601	0.505	0.549	0.579	0.497	205	136

*Sum Predicted and Misspecified columns represent total number of that class

Figure 10.6.1: Full evaluative metrics to assess predictive power of models using five-fold stratified cross-validation

Figure 10.6.1 demonstrates that random forest and decision tree algorithms are particularly robust models for injury prediction with high recall and F2-scores.

Figure 10.6.2: Full evaluative metrics assessing performance of models, tuned on recall scores, on test data								
Model	Class	Recall	Precision	F1-score	F2-score	AUC-ROC	Predicted	Misspecified
Baseline	No injury imminent	0.469	0.944	0.626	0.521	0.449	67	76
	Injury imminent	0.429	0.038	0.070	0.140	0.449	3	4
Custom OneR	No injury imminent	0.993	0.966	0.979	0.987	0.639	142	1
	Injury imminent	0.286	0.667	0.400	0.323	0.639	2	5
Decision Tree	No injury imminent	0.930	0.964	0.947	0.937	0.611	133	10
	Injury imminent	0.286	0.167	0.211	0.250	0.611	2	5
Random Forest	No injury imminent	0.993	0.959	0.976	0.986	0.593	142	1
	Injury imminent	0.143	0.500	0.222	0.167	0.593	1	6
Support Vector Machine	No injury imminent	0.000	0.000	0.000	0.000	0.500	0	143
	Injury imminent	1.000	0.047	0.089	0.197	0.500	7	0

*Sum Predicted and Misspecified columns represent total number of that class

Figure 10.6.2: Full evaluative metrics to assess predictive power of models using five-fold stratified cross-validation

However, Figure 10.6.2 shows all complex models performed poorly on test data with $F2 \leq 0.25$. Validation models outperformed them with OneR scoring $F2 = 0.323$. The test data results demonstrate the importance of choosing a suitable scoring metric to tune the hyperparameters of the models, with tuning on F2-score performing much better than on recall in the results section. As seen with the SVM results, tuning with certain metrics can cause neglect of other important aspects. This can have consequences down the line, such as generation of unnecessary prevention plans for users due to high share of non-injury risk individuals classified incorrectly.

Figure 10.6.3 illustrates the hyperparameter tuning for these models.

Figure 10.6.3: Machine Learning Models — Hyperparameter Tuning with Recall score			
Model	Parameters to tune	Parameter grid set	Optimal parameters
Baseline	—	—	—
OneR custom	Minimum samples to split at decision node	min_samples_split : [2, 4, 6, 8, 10]	4
	Minimum samples at each leaf node	min_samples_leaf : [1, 2, 3, 4, 5]	1
	Function to measure split quality	criterion : ['gini', 'entropy']	gini
Decision Tree	Maximum number of decision nodes	max_depth : [3, 5, 7, 10, 12, 20]	7
	Minimum samples to split at decision node	min_samples_split : [2, 4, 6, 8, 10]	2
	Minimum samples at each leaf node	min_samples_leaf : [1, 2, 3, 4, 5]	1
	Function to measure split quality	criterion : ['gini', 'entropy']	entropy
Random Forest	Number of decision trees constructed to vote on classification	n_estimators : [10, 20, 30, 50, 60, 75, 100]	75
	Maximum number of decision nodes per tree	max_depth : [3, 5, 7, 10, 12, 20]	20
	Minimum samples to split at decision node	min_samples_split : [2, 4, 6, 8, 10]	2
	Minimum samples at each leaf node	min_samples_leaf : [1, 2, 3, 4, 5]	1
	Random sampling with or without replacement when constructing decision trees	bootstrap : [True, False]	False
Support Vector Machine (SVM)	Regularisation parameter controlling aim to classify all points correctly at cost of more complex hyperplanes	C : [0.1, 0.5, 1, 1.5, 2]	0.1
	Degree of influence a single training instance has	gamma : ['scale', 'auto']	auto
	Defines the degree of the 'poly' kernel	degree : [2, 3, 4, 5]	3
	Defines the dimensional space and learning of the hyperplanes	kernel : ['linear', 'poly', 'rbf', 'sigmoid']	linear

Figure 10.6.3: Details of hyperparameter tuning settings, including parameter grids for each model

10.7 Glossary

Brief glossary of terms	
Term	Definition
Interaction facet	<p>Epidemiologically, Porta (2016) defined as <i>“the interdependent, reciprocal, or mutual operation, action, or effect of two or more factors to produce, prevent, control, mediate, or otherwise influence the occurrence of an event.”</i></p> <p>OR, also perceived as effect modification.</p>
Effect Modification	<p>Porta (2016) defined as <i>“Variation in the selected effect measure for the factor under study across levels of another factor”</i>. This implies when factors interact, the product may be greater (synergism) or lesser (antagonism) than either of the elements acting independently (Meeuwisse, 1994).</p>
Confounding facet	<p>Epidemiologically, Porta (2016) defined as: <i>“Loosely, the distortion of a measure of the effect of an exposure on an outcome due to the association of the exposure with other factors that influence the occurrence of the outcome. Confounding occurs when all or part of the apparent association between the exposure and the outcome is in fact accounted for by other variables that affect the outcome and are not themselves affected by exposure.”</i></p> <p>Meeuwisse (1994) simplifies this to <i>“An observed association between any two variables of interest could be due, totally or in part, to the effects of a third variable”</i>.</p>
Acute Workload	<p>The sum workload performed by an individual over a seven-day period. Can be interpreted as the ‘fatigue’ aspect of ACWR.</p>

Chronic Workload	Provides an indication of the training and activity level of an individual up to the present day. It is typically the four-week average acute workload. Viewed as indication of athlete fitness.
Workload	Measure of stress activity places on individual estimated by calculating the session RPR $sPRE = RPE \times duration$ of training session.
Acute Chronic Workload Ratio (ACWR)	Measured by the acute workload over the chronic workload, it provides a ratio of fatigue against fitness. It can help provide indications of how much an individual is over or under performing compared to what their body is used to. Hence, can be useful for understanding the preparedness of an individual and

10.8 ReadMe File



Injury Risk Prediction Modelling in Python

George Archer

MIT License

Copyright (c) 2024 George Archer (username:aca247)

Permission is hereby granted, free of charge, to any person obtaining a copy of this software and associated documentation files (the "Software"), to deal in the Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software, and to permit persons to whom the Software is furnished to do so, subject to the following conditions:

The above copyright notice and this permission notice shall be included in all copies or substantial portions of the Software.

THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.

Contents

Introduction.....	4
Setup instructions	4
Components.....	5
NHANES_data:	6
Data_flattening:	7
distributions:.....	8
Create_full_data_record_json:.....	9
sdv_metadata_generator:	11
synth_data_gen_sdv:.....	12
EDA_visualisation:	13
custom_snapshot:.....	13
ML_models:.....	14
Injury_predictor	16
Reference List:	17

Introduction

This file provides a breakdown of the set of classes and functions for the prediction of injury risk of users of Optimi Health app.

Objective

The aim of this set of programs is to construct a process which can take real world data, from Optimi Health, and return classifications of their app users' risk of injury via machine learning algorithms. Recognising the industry-felt shortage of data and general privacy issues, one of the stepping stones for building a predictive model has been a synthetic data generator to allow further data exploration.

Challenge

During the course of this project, the objectives had to be tailored and readjusted with delayed real data receipt affecting the scope achievable. As such, artificial data had to be constructed to inform and build the processes required.

Setup instructions

There exist a multitude of python dependencies to run all of the code. As such a requirements file exists. First set the working directory to the location of the scripts (cd "...\\Programs"). To ensure all the dependencies are installed, in the terminal input paste and run: *pip install -r requirements.txt*

Then run whichever file you want. You can use the sample artificial data file and sample models. Or, run file *create_full_data_record_json* to generate a new random artificial data file.

The python files *model_train* / *A_sample_on_running_here* provide an example of how a user can use the functionality, training selected or all models on a specified JSON dataset. They also demonstrate how the *injury_predictor* functionality can be used to predict injury for a specified dataset.

Components

To ensure accessibility across devices and potential users of the functionality outlined in this readme, the multi-step process was established in a virtual environment. This means that any packages required for operations of the functions are incumbent installations in the virtual environment.

Table 1.1 provides a brief summary of the key functionality.

Table 1.1 Virtual environment functions	
Class/Function	Description
NHANES_data	This class defines functionality to extract real world data from the NHANES database.
create_full_data_record_json	This class defines a set of functions that culminate to create an artificial dataset in place of the Optimi Health dataset.
data_flattening	This function transforms the input data from a JSON file into a more accessible pandas data frame format.
sdv_metadata_generator	Function generates metadata in the format appropriate for the synthetic data vault library, with input data having been passed through the <i>data_flattening</i> function.
synth_data_gen_sdv	Function takes the data output from <i>data_flattening</i> and respective metadata generated by <i>sdv_metadata_generator</i> to then create a synthetic dataset.
distributions	This provides a function which, given data inputs and distribution of a variable, randomly draws a value from desired distribution
EDA_visualisation	This defines a set of functions based on different chart formats, allowing easier consistent format chart creation. Also outputs and saves a set of charts to a designated folder – showing data characteristics.
Custom_snapshot	Takes a dataset with multiple records per user and creates a data snapshot such that all users appear once and the injury incidence meets a certain minimum
ML_models	Runs a set of machine learning models to predict injury
Injury_predict	Functionality for a user to run a trained model and predict injury

NHANES_data:

NHANES_data(construction = 0, file_path_demo = file_demo, file_path_bmx, file_path_d3 = None, file_path_d4 = None):

Overview:

The NHANES_data class is designed to extract data from the NHANES database and merge the different datasets to provide a full user data record.

Why needed:

This functionality was created to aid the development of more realistic artificial data. More specifically, the base functionality enables artificial data to be informed of the covariant relationships between personal statistics such as sex, height, mass and age – rather than random generation allowing for unrealistic height and mass combinations.

The process:

The NHANES 2017-18 data contains interview and examination data for 8704 persons residing in the United States, in order to provide high quality health data (NCHS, 2018b). The database splits datasets into multiple xpt files, meaning that certain characteristics are split among data files.

Class functions merge the two base data files "BMX_J.XPT" and "DEMO_J.XPT" on an ID variable, and keeps only relevant variables, notably age, sex, height and weight to construct personal statistics. (Note: The base data is collected for the period 2017-2018)

The functionality of the class also permits a user some personalisation, allowing them to define and merge other datasets from the NHANES database, and keep certain variables, at their own discretion (via d3 and d4). This is permissible via a binary input variable 'construction', where a value of 0 means the base dataset is extracted and created. Meanwhile a value of 1 allows for user customisation.

Dependencies: NHANES data files (NCHS, 2018a; 2018c; 2018d)

Data_flattening:

Data_flattening(data_file_path)

Overview:

The purpose of this function is to process the raw json data file provided by Optimi Health and to re-design the data structure from nested dictionaries to a pandas data frame format.

Why needed:

This functionality is needed to ensure that the data is in a format accessible to pre-defined python libraries, and to allow easier accessibility of aggregated data for inspection and visualisation. For example, the synthetic data vault library requires data in the format provided by the *data_flattening* function.

Process:

For each user in the raw json data input file, the functionality reallocates each user record and respective data value to among new data frames resulting in a dictionary containing the data frame names and actual data, i.e. `df_set = {'primary' : primary_df, 'training_load' : training_load_df,...}`

Dependencies: Inputted JSON data file in standard format

distributions:

distribution(low_limit, upper_limit, dist='right')

Overview:

This function is designed to make the random value assignment of data variables more meaningful rather than assuming a more simplistic distribution.

Why needed:

As a result of real data not surfacing this has necessitated the creation of some form of data in order to construct subsequent data processes and retrieve an output of some kind. Hence, given paucity of similar data online, some of the specific variables have to be generated via random assignment between bounds or from a set list. Given the variety of data values, data variables likely possess different distributions, for example the RPE/intensity measure on a 0-10 scale likely requires a left-skew distribution given that the nature of training involves physical exertion and few likely scoring in the 0-2 range.

Process:

The function requires 3 inputs, the bounds [low_limit, upper_limit, and the type of distribution the data variable is believed to be characterised. There are five different options for the distribution:

- 'right' : right-skewed distribution
- 'left': left-skewed distribution
- 'uniform': uniform distribution
- 'symmetric': symmetric distribution
- 'dec_exp': decreasing exponential distribution

The majority of these are constructed via use of the Beta distribution and different specifications of alpha and beta parameters to design a generic shape of the desired distribution. In this way, can assign random values in a slightly more realistic manner. For instance, it is common belief that the share of the population with x number of injuries exponentially decreases as x increases, hence use if 'dec_exp'.

This functionality is instrumental in generating artificial data within the *create_full_data_record_json* class.

Create full data record json:

`Create_full_data_record_json(rand_seed=0, dist=0)`

Overview:

This class is designed to bundle together data and functionality which work together to create different instances of artificial data, based off a random seed assignment.

Why needed:

As a result of real data not surfacing this has necessitated the creation of some form of data in order to construct and evaluate subsequent data processes as well as attain an output of some kind. Additionally, due to the nature of certain variables not being mutually exclusive, and distributions not being uniform, data creation has to be thoughtfully crafted to ensure some inference rather than just full random generation.

Process:

The class has multiple functions in order to produce the end result of informed artificial data matching the style of Optimi Health data.

Initiation(rand_seed=0, dist=0):

- Initiates the class calling the default dataset from the NHANES data involving a set of records regarding age, weight, height and gender
- Establishes the random seed and whether using distribution properties or actual data for primary_stats, dist=0 means just attach other generated data to the real NHANES data subset while dist=1 means attain distribution properties from NHANES data and attach.

create_demo_stats(user = None):

- Used if dist = 1 to review distribution of age, sex, weight and height and then assign values to a user based on these distributions and interdependent relations

create_training_load_acwr(user = None, signup_date = None, training_practice = {'Cycling': 2, 'Tennis': 2, 'Hiking': 1}, data_format = 'split'):

- Takes a randomly generated list of routine activities and their respective weekly frequency for a single user, in the *training_practice* input, and user sign up date to the app to create a time series record of the user's training routine since use of the app
- Assumes that for each day the user has a set of routine/fixed exercises they perform each week i.e. have football / cricket match play on Saturdays or practice/lessons on Tuesdays every week.
- From this generates a historical record of activity with randomly generated exercise duration and RPE / intensity where the RPE is assumed to follow a left-skewed distribution. RPE varies per week, while session duration is assumed constant week-to-week due to routine nature. These values allow the calculation of the acute-chronic workload ratio (ACWR)

- To add additional randomness a binary determinator has been added to determine if the user is doing the exercises or not. This 'prob_routine' is to account for sickness, other activities and general life events.
- Similarly, there is the potential to incorporate a "spontaneous activity", i.e. friends want to go kayaking this weekend. For now, this has not been included but can be applied in a similar fashion.
- The function returns two datasets, either as individual datasets if 'data_format' is specified as 'split' or as a singular dataset if specified as 'joined'. The 'split' format availability is due to the proposed data structure separating training load and ACWR data.

create_injury_history(user=None, signup_date = None, age = None):

- This randomly generates a user's injury history, specifying a body part from a random uniform distribution and specifying time considerate injury start and end dates.

create_wearable_data(user=None):

- Uses academic study analysis of resting heart rate to provide distributional data (Galarnyk et al., 2020)
- Uses data from NHIS survey to randomly select from for sleep duration (NCHS, 2018a)

create_forms_of_training(user, number_of_methods = 2, sessions_pw=2):

- Randomly selects sample of physical activity methods (based off number_of_methods input, which is determined by the user type) and assigns frequency per week of this routine activity given user type bounds on minimum and maximum number of sessions a week.

create_user_data(num_users=10):

- Collates all data to generate records for the number of user (num_users) specified in the function input.
- Generates user ids and iterates through the users, assigning a user type (routine, specialist or infrequent) which determines some base user characteristics to then iterate through the prior functions mentioned
- Returns a dictionary of the pandas dataframes containing all artificial data

preprocess_data_for_json(data):

- Important to ensure the artificially generated data from *create_user_data()* function is in JSON format with datatypes being transferred to data formats

export_rand_data_as_json(num_participants = 500, filename = 'artificial_injury_data_sample'):

- Utilises the *create_user_data()* function to create a dataset with *num_participants* users and respective records
- It then formats the artificial dataset using the *preprocess_data_for_json()* function and exports the JSON to the respective *filename* in the working directory.

sdv_metadata_generator:

`sdv_metadata_generator(flattened_data_dict)`

Overview:

This function aims to generate the metadata from the raw data file, following appropriate formatting done by the `data_flattening()` function.

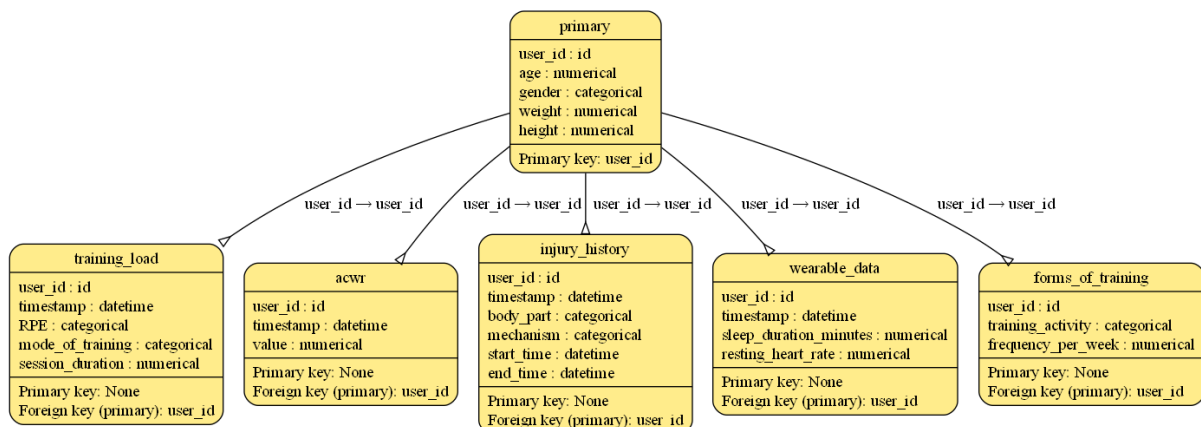
Why needed:

In order to utilise the python library *Synthetic Data Vault (SDV)*, and run a synthetic data generator for a multi-table relational dataset, two inputs are needed: the metadata describing each table and the relationships, and secondly the dataset in the format of a dictionary mapping table names to dataframes (Datacebo, 2023a)

This function generates the metadata format, and specifies the library specific data type, known as sdtypes, for the variables.

Process:

The python library provides its own functionality to help with the process, automatically translating the incumbent data types to its own data types. Creating an instance of the MultiTableMetadata class and using the “detect_from_dataframes” function performs this for its users (Datacebo, 2023b). Some manual adjustments such as ensuring ID variables are classified as such are required, and to aid understanding a graphic illustrating relationships is constructed, as below, and saved in the working directory, while a JSON of the metadata is also exported.



[synth_data_gen_sdv:](#)

`synth_data_gen_sdv(dataset_file_path)`

Overview:

This function aims to take a JSON file as an input and return a synthetic dataset, anonymising the data and scaling the dataset allowing more records to be generated.

Why needed:

This functionality is useful in the modelling space as it allows the creation of realistic data. This has potential to allow greater deep learning models and the construction of neural networks. Hence, the synthetic data generator can permit the creation of more complex models with higher predictive power and greater accuracy.

Process:

There are three main steps to the synthetic data generator, however a preliminary step is to process the input file into an appropriate dataframe format via the `data_flattening()` function and generate the respective metadata via the `sdv_metadata_generator()` function.

The base steps are:

- 1) Creation/initiation of the synthesiser using the metadata
- 2) Training of the synthesiser using the "real" data
- 3) And finally, the generation of the synthetic data based on defined parameters

The synthetic data vault library provided an appropriate means for injury prediction due to its ability to accommodate relational databases via its multitable functionality. The library has four synthesisers available but only the Hierarchical Modelling Algorithm (HMA) synthesiser is freely available (Dacebo, 2023c).

EDA visualisation:

Overview:

This function aims to provide data visualisations of the whole dataset.

Why needed:

This functionality helps an individual understand the dataset, especially the artificial dataset providing visual access to key metrics and distributions of data.

custom_snapshot:

Overview:

Due to significant class imbalances with injury positive instances being a minority snapshot analysis was not plausible due to insufficient instances for ADASYN to generate additional records. Functionality produces a custom snapshot by collating data from other sources.

Why needed:

This functionality is essential to ensure enough positive instances exist in the dataset to train the machine learning models.

ML_models:

ML_models(dataset = input_data, outcome_var = 'injury_imminent', min_instances = 400, min_positive_instances = 0.04, rand_seed_for_snapshot_gen = 0, directory=save_file_path)

Overview:

The script is designed to train and evaluate a variety of machine learning models to predict the injury imminent classification of data instances. The models include a baseline, custom OneR, decision tree, random forest and support vector machine.

The script offers flexibility in model selection and utilises hyperparameter tuning, cost matrix and stratified cross validation to improve the models and provide sufficient evaluation via its custom metrics. To address data paucity and class imbalances, it automatically applies ADASYN and provides synthetic data generation options.

It automatically creates directories to save details on the data analysed, models predicted, including graphics and evaluation metrics with appropriate timestamps.

Why needed:

The script provides the key analysis for the whole project for which all prior scripts detailed have been produced for and predicts the injury risk of individuals.

Process:

The class has a set of functions defined within it:

Function	Description
snapshots_suitable()	Assesses if any naturally-occurring data windows contain sufficient number of positive cases
data_preprocessing(data_snapshot, vars_to_remove, test)	Simply processes the data, separating x and y variables, splitting data into train and test sets, and applying ADASYN to the training set only
abstract_dataset(injury_incidence, rand_seed)	If no naturally-occurring data windows according to snapshots_suitable, abstract_dataset treats time as irrelevant and generates a data window by pulling data for each user from across the dataset – ensuring there is a minimum injury_incidence ratio
eval_metric(y_test, y_pred, y_pred_proba, *, class_names)	Functionality to return the evaluation metrics for predictive models. It returns a host of metrics, namely recall, precision, f1-score, f2-score, AUC-ROC and specification numbers.
Scv()	Conducts split number of fold stratified cross validation to properly assess the model's predictive power.

x_train, y_train, model, splits = 5, rand_state = 7)	
hp_tuning_gs(model, x_train, y_train, param_grid, num_folds_cv = 5, scoring = 'f2')	For inputted model type, with predefined model-specific parameter grid, the function finds the optimal set of hyperparameters to maximise the recall score of the model. It does this by using the experimental HalvingGridSearchCV to minimise computational resources while sustaining likelihood of finding the optimal parameters.
baseline	Uses stratified technique to train a model
oner_custom	Based off the OneR algorithm, single level decision tree
decision_tree	Hyperparameter tuned decision tree
random_forest	Ensemble method, tuned
svm	
desc_stats	Generates some base descriptive statistics
model	Allows the choice of model to be run, or all, trains and creates subsequent graphics and saves all details to created directory
Predict(trained_model, raw_json_file, evaluate=0)	Takes the trained model in form of a joblib file, and using this predicts the injury risk of instances in the raw_json_file

Most of the functionality is pre-defined, to see how to run code and experiment see the example functionality:

```
if __name__ == "__main__":
    root_dir = os.path.dirname(os.path.dirname(os.getcwd()))
    data_wkd = os.path.join(root_dir, 'Artificial data')
    filename = "artificial_injury_data_sample.json"
    filepath = os.path.join(data_wkd, filename)

    input_data = single_data_record(filepath, use_synth_data=0) # increase dataset so at least 1000 records

    test = ML_models(dataset=input_data, outcome_var='injury_imminent')

    test.model('all', 'custom', hypertune = 1, class_weights_var='b')
```

The above code uses the input data from *filepath*, it then processes this data allowing for the generation of synthetic data. Then *ML_models* is initiated and assigned to *test* identifying the outcome variable for prediction as 'injury_imminent'.

Then all models are run on the data, using a custom data snapshot, ensuring the models' hyperparameters are tuned and class weights balanced. With the use of grid searching this process of generating the optimal models and their parameters can take over two hours – specifically due to the support vector machine (other models trained faster).

Injury_predictor

Overview:

This function aims to take a JSON file as an input and using a trained model predict the injury risk of each user.

Why needed:

This functionality is useful in providing an easily accessible for the non-creator to use.

Reference List:

1. National Center for Health Statistics (NCHS), 2018a. *National Health Interview Survey (NHIS) 2017 Data Release: Sample Adult file, CSV data* [Online]. National Center for Health Statistics. Available from: https://www.cdc.gov/nchs/nhis/nhis_2017_data_release.htm [Accessed 23 August 2024].
2. National Center for Health Statistics (NCHS), 2018b. *National Health and Nutrition Examination Survey (NHANES) 2017-2018* [Online]. National Center for Health Statistics. Available from: <https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?BeginYear=2017> [Accessed 14 August 2024].
3. National Center for Health Statistics (NCHS), 2020a. *National Health and Nutrition Examination Survey (NHANES) 2017-2018 Data Documentation, Codebook, and Frequencies – Demographic Variables and Sample Weights (DEMO_J)* [Online]. National Center for Health Statistics. Available from: https://wwwn.cdc.gov/Nchs/Nhanes/2017-2018/DEMO_J.htm [Accessed 14 August 2024].
4. National Center for Health Statistics (NCHS), 2020b. *National Health and Nutrition Examination Survey (NHANES) 2017-2018 Data Documentation, Codebook, and Frequencies – Body Measures (BMX_J)* [Online]. National Center for Health Statistics. Available from: https://wwwn.cdc.gov/Nchs/Nhanes/2017-2018/BMX_J.htm [Accessed 14 August 2024].
5. National Center for Health Statistics (NCHS), 2018a. *National Health Interview Survey (NHIS) 2017 Data Release: Sample Adult file, CSV data* [Online]. National Center for Health Statistics. Available from: https://www.cdc.gov/nchs/nhis/nhis_2017_data_release.htm [Accessed 23 August 2024].
6. Galarnyk, M., Gouda, P., Quer, G., Steinhubi, S.R. and Topol, E.J., 2020. Inter- and intraindividual variability in daily resting heart rate and its associations with age, sex, sleep, BMI, and time of year: Retrospective, longitudinal cohort study of 92,457 adults. *PLOS One* [Online], 15(2). Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7001906/> [Accessed 23 August 2024].
7. Datacebo, 2023a. *Synthetic Data Vault: Multi-table-data – Data Preparation* [Online]. Available from: <https://docs.sdv.dev/sdv/multi-table-data/data-preparation> [Accessed 18 August 2024].
8. Datacebo, 2023a. *Synthetic Data Vault: Multi-table-data – Data Preparation Metadata API* [Online]. Available from: <https://docs.sdv.dev/sdv/multi-table-data/data-preparation/multi-table-metadata-api> [Accessed 18 August 2024].

9. Datacebo, 2023c. *Synthetic Data Vault: Synthesizers* [Online]. Available from: <https://docs.sdv.dev/sdv/multi-table-data/modeling/synthesizers> [Accessed 18 August 2024].