Evaluating predictive models of recreational water quality at Auckland beaches

Jill Bolland – jbol001 Linda Hu – nhu002 Vishal Thamizharasan - vtha709 Channing Wang – cwan299

***Abstract*—Auckland Council has recently set up a Safe Swim website with a daily rating of water safety at popular swimming beaches. The safety measure is based on the amount of faecal bacteria in the water. The most accurate way to measure water quality is sampling and laboratory testing of bacterial levels at each site but this can take several days so that the information is out of date and not useful for daily decision making. To address this models that can predict bacteria levels in advance have been developed. Our research will build on recent work to build machine learning models for water safety prediction and aims to improve prediction accuracy by including additional features, comparing sampling techniques for unbalanced data, and addressing temporal issues in the sampling data.**

***Index Terms*—recreational water quality, adaptive synthetic sampling algorithm, machine learning algorithms**

I. Introduction

The ability to safely swim at a beach is seen as a right of passage for most New Zealanders. A 2014 Safe Swim survey by Auckland Council found that beaches and harbours were the most valued part of the environment and that beach water quality was a key concern. Auckland Council has responded to this sentiment by monitoring bacteria levels at Auckland's most popular swimming beaches and providing up to date information on Safe Swim. A beach is deemed safe to swim if sampled bacterial counts are below a set threshold. Unfortunately it takes around 48 hours to analyse water samples and by this time the data is out of date and so cannot be used by the public when deciding on whether to engage in water activities. Therefore, the council commissioned a model which can predict the presence of dangerous levels of bacteria in advance. The model predicts whether Feacal indicator Bacterial (FIB) levels will exceed levels deemed safe for swimming.

Many jurisdictions provide water quality indicators to inform the public and officials on the safety of water in their regions. In 2013 the United States Geological Survey [1] recommended standards for water safety model prediction accuracy. The guidelines suggest using other performance measures along with overall accuracy as this can be misleading when water quality is normally high and there are relatively few occasions when bacterial levels are unsafe. In these situations, sensitivity is a better metric as it measures the proportion of correctly identified dangerous cases. Specificity measures the ability to detect a negative case based on the number of correct negative prediction over the total number of negative samples. The USGS guidelines recommend a model must have accuracy over 85%, sensitivity at 50% and specificity at 80%. For a water quality predictive model to be included into the safeswim platform, it must meet the USGS guidelines [1].

Much of the data used in modelling water quality is very unbalanced as high levels of bacterial are unusual. In in our study region, Auckland, most of the samples have FIB levels below the health and safety danger threshold, while only a very small portion of the data are above the threshold. If the imbalanced dataset is not handled properly, any machine learning model will learn to predict the majority class which will result in high overall accuracy rate but poor prediction accuracy on occasions when the FIB levels are above the threshold. There are various mitigation strategies for handling unbalanced datasets for machine learning and a larger discussion of these methods are provided in Section II. An additional issue in this domain is that sampling data at a site over a time period is essentially time series data. If this data is changing over time then the entire modelling process must take this into account. Sampling and model validation and evaluation techniques have been designed specifically for time series. However, studies on water quality do not always include these techniques. Therefore this study will investigate ways of maintaining any similarity of samples due to their closeness in time.

Researchers in the Auckland University of Environmental Sciences Xu et al. [2] have worked with Auckland Council to assess the ability of machine learning techniques to predict water quality. Their research indicates that machine learning models may outperform the water quality prediction model currently used. The author also suggested two other improvements. Firstly incorporating other weather or site features which could improve forecasting and secondly the development of a more general model which could be used for all sites in the Auckland region.

The current study aims to extend the Xu et al. [2] analysis by including new features and building a model for all five beaches. We also incorporate time series techniques in the modelling and evaluation process.

II. Related Work

A. Predictive models of recreational water quality

Xu et al. [2] used water quality monitoring data from 1995 to 2018 for five beaches in Auckland. Each dataset had eight variables: Feacal indicator Bacterial (FIB), four rainfall accumulations for the past 24, 48, and 72 hours, and total accumulated precipitation, wind direction, wind speed, and solar hours per day. To address the problem caused

by imbalanced dataset, Xu et al. used an adaptive synthetic sampling algorithm (ADASYN) to generate synthetic data in minority class to make the training data more balanced [2]. ADASYN is an improved version based on Synthetic Minority Over-sampling Technique (SMOTE). The key idea is to generate synthetic samples using the density distribution of the minority real samples.

In Xu et al. [2] the FIB sample data were reclassified into two classes; negative where FIB levels were below 280 Colony Forming Units (CFU) per 100ml and positive where FIB levels were above 280 CFU per 100ml. These levels were decided based on New Zealand Guidelines. Four different algorithms: k-nearest neighbors algorithm (K-NN), boosting decision trees (BDT), support vector machine (SVM), artificial neural network (ANN) were trained to predict whether the FIB was above or below the threshold. Each algorithm used four different training datasets for each of the 5 beaches: the original unbalanced dataset, a balanced dataset built by under sampling the majority class, a balanced dataset built by over-sampling the minority class which included duplicate minority samples, and an ADASYN method generated balanced dataset. The experiment results suggested that the use of ADASYN generated balanced datasets gave the biggest improvement in model performance, particulary the ability to correctly classify positive samples.

While the ADASYN sampling method improved results, the authors suggested several areas for improvement including addressing overfitting problems caused by ADASYN generated samples which were too similar to majority class samples. They noted that the models were able to predict well for a single site but not a general model for all sites and that including beach characteristics as a feature may allow a more general model. In addition, in their model design, the testing dataset was not fully independent of the training dataset so model accuracy may have been overestimated.

In a study that did separate training and testing data prior to modelling, Zhang et al. [3] modelled fecal coliform (FC) bacteria levels at Holy Beach on the Louisiana Gulf Coast, USA [3]. They tested two predictive models an artificial neural network (ANN) and multiple linear regression (MLR) for predicting the FC bacteria level at Holy Beach. The data was collected from six sampling sites on the Holly Beach every Monday morning between May 1 and October 31 from 2005 to 2010. The data had 14 parameters: salinity, water temperature, wind speed (six categories from calm to strong), wind direction (off/on shore), individual tide effect (nine categories from extremely low to extremely high), cumulative tide effect, tide water level, weather type (sunny=1, cloudy=0), and 9 – 14), and six different antecedent rainfall parameters – one day before, two days before, three days before, rainfall accumulated in the last 48 hours, 72 hours, and 96 hours. The two models were then tested against unused datasets from 2005, 2006 and 2010. The measure of performance was the linear correlation coefficient (LCC). The ANN model outperformed the MLR model with a LCC of 0.437 compared to 0.120 for the MLR. While Zhang et al. [3] study did not have the issues of dependent testing data they did not address the imbalance in the dataset.

A study [1] evaluating multiple predictive models for beach management at a freshwater beach at Sandpoint Beach in the Great Lakes region of Canada, analysed water quality data collected over 5 years (2014-2018) during summer months. Water samples were collected at 5 locations from Sandpoint Beach at least once a week on Mondays from June to September. If the geometric mean of the E. coli counts was smaller than 200 CFU per 100ml, the health risk was deemed low and the water safe for swimmers. E.coli counts between 200 CFU and 999 CFU per 100ml indicates a potential risk and swimming in the water is not recommended. If the result is greater than 1000 CFU per 100ml, the beach is closed and water samples are recollected on Thursday of the same week. Modelled features included parameters about water quality and beach condition - number of birds; day of year; turbidity; water temperature at the time of sampling; and meteorological data including accumulated rainfall, wind direction relative to the shoreline, wind speed, daily air temperature and wave height. The models trained were MLR baseline models with only E. coli data, MLR models of water quality and beach condition but excluding qualitative weather information, and an MLR models using all of the parameters listed above. All the models used 2018 data as testing data, while compared on different training periods of 2-years, 3-years and 4-years respectively. Model performance was evaluated by estimating the area under the receiver operating characteristic curve (AUROC), This allows evaluation of the true-positive rate (sensitivity) against false-positive rate (1-specificity) rather than just an overall accuracy score. The study found that including qualitative weather data improved the MLR models' performance. MLR models generated with the 2-year training data performed better than the models generated with 3-year and 4-year training data suggesting that models may need be updated over time. The authors also suggested that it may be useful to analysis other beaches to determine how many years data is adequate for training the models.

B. SAMPLING

Handling imbalanced data, for classification, regression is an important issue as most of the traditional machine learning methods and evaluation metrics assume a balanced class distribution. Class imbalance occurs when there are far fewer samples for the minority class(s) than the majority class(s), as a result the learner algorithms will over-classify the majority class due to their increased probability while misclassifying the minority classes more frequently.

Seiffert et al. [4] noted that the number of samples for the minority class is more important than the percentage of the minority class for an imbalance dataset problem. The classifier can train on the dataset without compromising performance if there is a sizable number of minority classes. The majority class(s) is under sampled before training. Similarly, Krawczyk [5] noted that as long as both majority and minority class(s) are well represented and come from non-overlapping distributions,

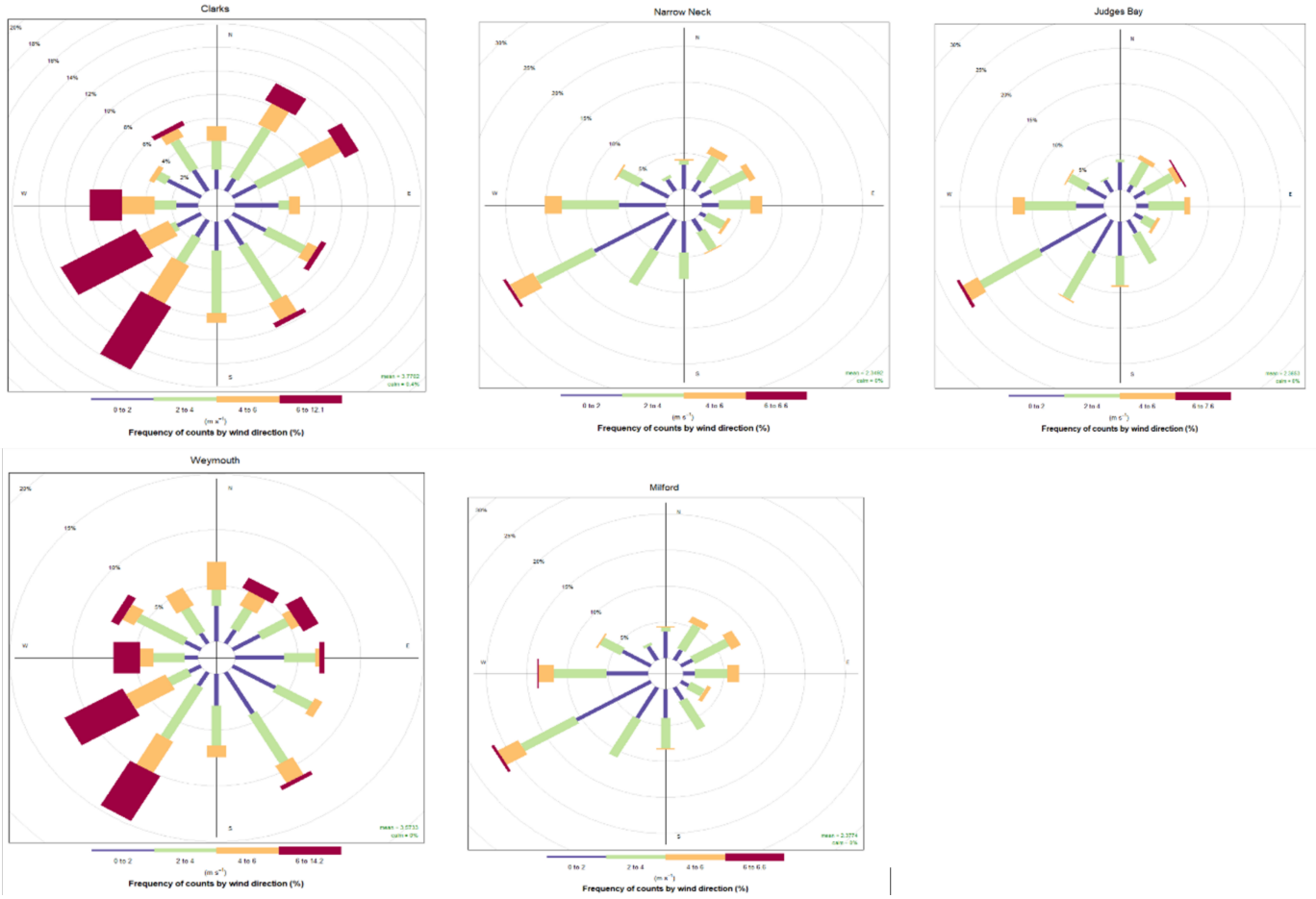


Fig. 1. Wind speed by Wind direction for each Beach

standard classifiers can produce good classification rates. This is not the case with our water quality data set where there are only 149 minority class instances and both classes come from overlapping distributions. With a class imbalance ratio p = 12.5, falling within the guideline set in Krawczyk [5] (imbalance ratio ranging from 4 to 100, where p = ratio of majority class sample number by minority class sample number), oversampling techniques can be used for the Auckland dataset.

Oversampling techniques can range from classical resampling methods like synthetic minority oversampling technique (SMOTE) and its variants (ADASYN, Safe-level SMOTE, Border-line SMOTE) to newer methods that consider the overall minority class distribution like RACOG (RApidy COnverging Gibbs), wRACOG (Wrapper-based RApidy COnverging Gibbs), and RWO-Sampling (Random Walk Over-Sampling) as noted in Kong et al. [6].

To benchmark the various sampling approaches, Kong et al. [6] tested three SMOTE variant sampling approaches versus the newer sampling approaches, RACOG, wRACOG, and RWO-Sampling on 19 benchmark datasets that have imbalance from the KEEL-collection [7] (classification task on E. coli type, yeast type, glass and vehicle dataset). They found that the newer methods outperformed the classical approaches largely, with RACOG performing the best.

The oversampling techniques listed above cannot be directly applied to a time series data as they do not take temporal information of the data instance into account. The SMOTE algorithm and its variants randomly choose two parent minority class instances to generate a synthetic offspring, in terms of a time series dataset, the application of the original oversampling algorithm can include parents from different temporal regions.

Cao et al. [8] proposed a multi-step methodology to generate minority class samples from an imbalanced time series dataset via structure preservation. First the null space is removed and each of the training time series vectors is transformed into a low dimensional signal space. The advantages include better oversampling in a lower dimensional space and elimination of the potential risk an oversampling method artificially introduces data variance in the common null space, which originally contains no data variance. Next, estimation of the minority class covariance and regularization of the eigen spectrum is performed, on this oversampling is done by preserving regularized positive covariance. These are labelled as enhanced structure preserving oversampling (ESPO) samples. Finally, ADASYN is applied to the ESPO samples and the

original minority class samples to generate the samples that will emphasize (protect) the existing minority instances close to the classification boundary. An implementation of this methodology exists in the OSTSC R package Dixon et al. [9] The authors have demonstrated significant improvement in AUC scores for various time series datasets that have been oversampled using the integrated ESPO+ADASYN (INOS) approach.

C. Model Evaluation

When dealing with extremely imbalanced dataset, evaluating machine learning algorithms with predictive accuracy is not appropriate. For example, a normal mammography dataset may only have 2% abnormal pixels with all the rest are normal. If a model only predicts normal for every case, it will have a predictive accuracy of 98%, which loss the significance of predicting the abnormal cases. The Receiver Operating Characteristic (ROC) curve is a better approach for evaluating predictive model with imbalanced dataset as it compares the true positive rate and the false positive rate. The area under the ROC curve is a useful metric for comparing the performance of different models. Auckland Council uses international standards of accuracy for water quality models and any model used must have at 85% accuracy, 50% sensitivity, and 80% specificity scores. Therefore, we will evaluate our models using these measures.

Given the large variety, constantly growing, number of machine learning algorithms now available as open source and accessible to all, researchers require a way to statistically compare their predictive performance. Looking at studies published in proceedings from the International Conference on Machine Learning in over 4 years, Demsar [10] concludes that there are currently no agreed and standard tests to compare multiple classifiers. T-tests have traditionally been used but these only compare two classifiers on a single dataset and these parametric tests make strict assumptions around data normality and equality of variance. Unfortunately it is unlikely that all classifiers will produce a similar distribution of performance scores. In addition if more than two classifiers are compared care needs to be taken as repeating t-tests increases the risk of rejecting the null hypothesis, of no difference in classifier performance, when it is true. This is because the tests use the probability of finding a difference just by chance in their evaluation of the test statistic. While critical values can be lowered or factors applied to account for multiple testing, the authors recommend instead that machine learning practitioners draw from statistical theory and utilise tests specifically designed for multiple comparisons. These include the parametric ANOVA and the non-parametric Freidman Test.

The current study aimed to build and evaluate several machine learning algorithms to predict whether it is safe to swim at five beaches in Auckland. Models were built for each individual site along with a more general model that can be used for all of the five beaches. Extra features including beach type (coastal or tidal) and wind direction relative to the shoreline were included in addition to the original features included in Xu et al. [2] As initial data exploration indicated that Entero levels were increasing over time we incorporated time series in our model cross validation and final evaluation SMOTE sampling, designed to handle categorical data was used to generate a more balanced training set. We compare 4 machine learning algorithms trained on data from individual beach sites and data from all sites pooled. A recursive feature elimination (RFE) method was used to determine the best features to include . We also fit a baseline multiple logistic regression model to the original data and to oversampled data to demonstrate the expected improvement in model performance that can be achieved by generating a more balanced dataset. The final model results were assessed using tests statistical tests designed to discriminate between multiple models.

III. Methodology[1]

A. DATASET

We obtained the original data from M. Neale, one of the authors of the research paper [2]. The data is water quality monitoring data from 1995 to 2018 for five beaches in Auckland, including Clarks Beach, Narrow Neck Beach, Judges Bay Beach, Weymouth Beach, and Milford Beach. Each dataset has nine variables: date the sample was collected, Entero counts, rainfall amount accumulated during the past 24, 48, and 72 hours, and total accumulated precipitation, wind direction, wind speed, and solar hours per day. With all beaches combined there were 2,017 observations that had the full set of features. Figure 2 shows that numbers of observations across years are inconsistent, the lowest amount of samples was 9 during 1995 and the highest was 140 during 2001. After consultation with Auckland Council, we ascertained that there was a budget for sampling work and this budget varied over time and so the number of samples taken also varied. There is also evidence that the volumes of Entero in the sampling data has increased over time. This can be seen in Figure 3 which shows Entero counts from Narrow Neck beach during the sampling period, therefore there is likely to be some time signals in the data. Further consultation with a laboratory technician indicated that sampling methods also changed during over the data time period.

As per Xu et al. [2] we reclassified the FIB sample data into two classes. If the Entero counts were below 280 colony Forming Units (CFU) per 100ml, the FIB levels were classed as negative; and classed as positive where Entero counts were above 280 CFU per 100ml. These levels were decided based on New Zealand Guidelines [2]. Figure 4 shows distributions for each variable along with the paired scatter plots. The safe levels of Entero (negative samples) are represented by the blue dots and the unsafe level (positive samples) are represented by the orange dots. The scatterplots clearly show the imbalance in the data. The rainfall data in particular are highly positively skewed with many days of little or no rainfall and a small amount of days with large accumulated amounts.

[1] https://github.com/ChanningW2211/PB5

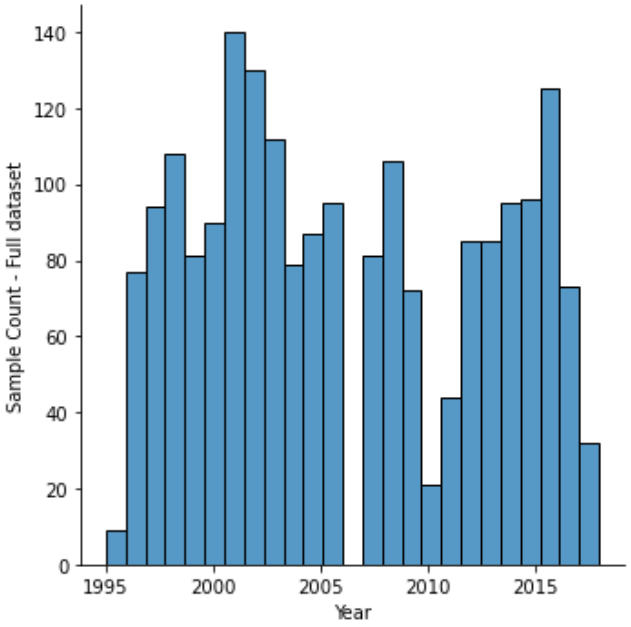


Fig. 2. Sample Count by Year

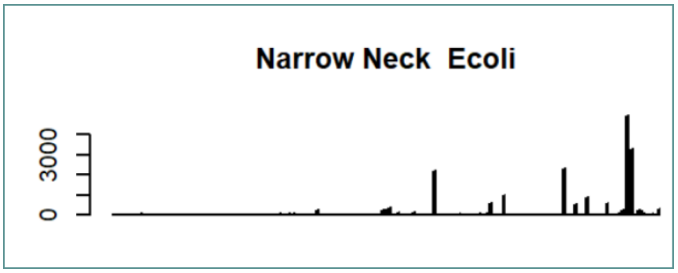


Fig. 3. Scatter plot by pairs

This is typical of rainfall data which generally follows a gamma distribution. To combat the skewness all variables were normalised using the min/max method before training the machine learning models.

B. FEATURE ENGINEERING

As suggested from previous research [3] site characteristics for the beaches where the samples were collected were included as variables. Firstly, beach type was added with two categories: open coast beach and sheltered bay and harbour. Milford and Narrow Neck are open coast beaches, while Clarks, Judges Bay and Weymouth are sheltered bay and harbour. According to our conversations with Xu et al. [2] open coast beach are thought to be less likely to accumulate bacteria as the tide can more easily flush the bacteria out to sea.

Figure 1 shows that the wind speed and direction by each beach is quite different so the wind direction in relative to the beach was also included.

The wind directions in the dataset were converted from 360 degrees to 16 cardinal directions [11]. Then beach directions were added by drawing perpendicular line from the coast

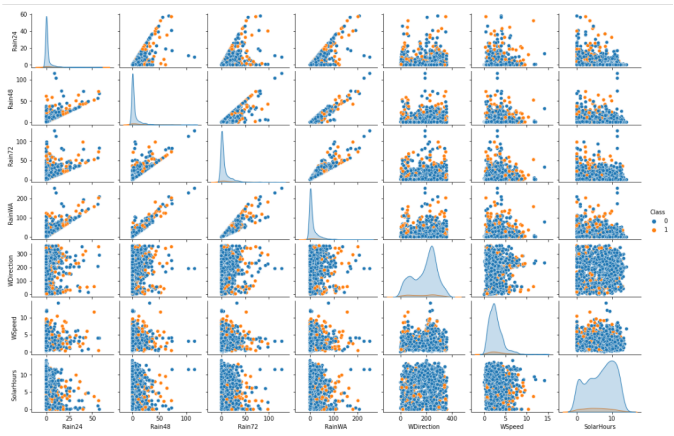


Fig. 4. Scatter plot by pairs

line of each beach to decide the cardinal direction of each beach. Then the on/off shore wind feature was generated by comparing the difference between the wind direction and beach direction. As demonstrated in Figure 5, since wind direction is generally reported by the direction from which it originates [12], wind directions in the green area are on shore and yellow area are off shore. The white areas are cross shore.

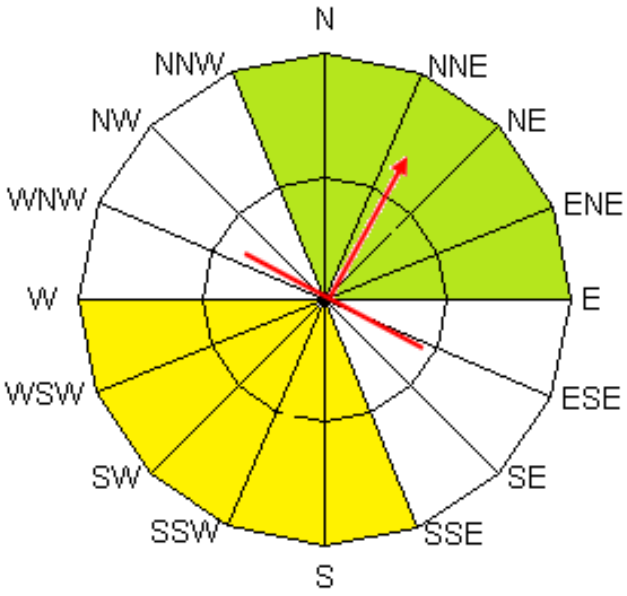


Fig. 5. On/Off Shore Feature Generation

As the total accumulated precipitation RainWA was equal to the sum of Rain24, Rain48 and Rain72 we removed this as it does not add any further information to the data [13]. Although there was a high correlation between accumulated rainfall amounts, not surprisingly, we decided to leave all of these variable in the data set as we could not accuratel determine which one was most important and it may be that sequential days of high rainfall may also affect bacteria levels. We used recursive feature elimination (RFE) methodology to

check which features should be included in the final models. RFE is a wrapper method which can use several machine learning algorithms to evaluate its performance [14]. The basic idea of wrapper method is demonstrated in Figure 6. Based on the RFE results we included the following features: Accumulated rainfall over the previous 24, 48, and 72 hours, wind direction and speed, Solarhours, BeachType, on_offshore wind direction.

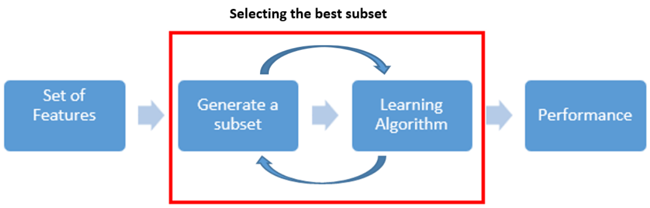
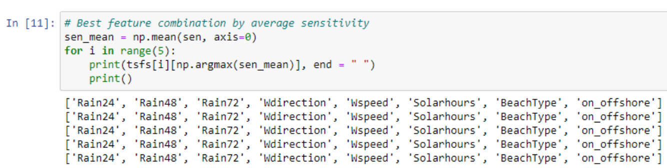


Fig. 6. Flow of Wrapper Method



```
In [11]: # Best feature combination by average sensitivity
sen_mean = np.mean(sen, axis=0)
for i in range(5):
    print(tsfs[i][np.argmax(sen_mean)], end = " ")
    print()

['Rain24', 'Rain48', 'Rain72', 'Wdirection', 'Wspeed', 'Solarhours', 'BeachType', 'on_offshore']
['Rain24', 'Rain48', 'Rain72', 'Wdirection', 'Wspeed', 'Solarhours', 'BeachType', 'on_offshore']
['Rain24', 'Rain48', 'Rain72', 'Wdirection', 'Wspeed', 'Solarhours', 'BeachType', 'on_offshore']
['Rain24', 'Rain48', 'Rain72', 'Wdirection', 'Wspeed', 'Solarhours', 'BeachType', 'on_offshore']
['Rain24', 'Rain48', 'Rain72', 'Wdirection', 'Wspeed', 'Solarhours', 'BeachType', 'on_offshore']
```

Fig. 7. Feature Selection Results

C. SAMPLING

As previously mentioned a solution to address the general problem of an imbalanced dataset is to use re-sampling techniques. These techniques change the distribution of the dataset by balancing the number of rare classes to majority classes, in effect reducing the skewness of the dataset. The common approach to re-sampling is either over-sampling or under-sampling the dataset.

The seed paper Xu et al. that we are extending upon on has shown that under-sampling significantly reduces the prediction accuracy for the below threshold samples, ie; the majority class. This could be due to severe under representation of the majority class samples that could potentially lead to loss of useful information. The majority class samples could have more variability in their features across the dataset and thus a random selection of majority class samples could miss some of this variation. Similarly the authors have shown that applying a standard oversampling on the minority class samples led to poor prediction accuracy for minority classes.

As previously discussed the authors have shown ADASYN (Adaptive Synthetic Sampling algorithm) to best over sample the dataset while attaining high sensitivity and specificity scores.

As part of our replication of the seed paper's implementation as our baseline model, we have also found ADASYN to give the best results in terms of accuracy, sensitivity and specificity when compared to SMOTE variants like Borderline-SMOTE

Technique	Description
ADASYN	It extends SMOTE by oversampling harder to learn samples, while adding a random value to the synthetic sample on the interpolated line to make it less linearly correlated with the parent samples.
Borderline-SMOTE	An extension to SMOTE, it identifies borderline samples and use these to generate the new synthetic samples. Samples on the borderline are more likely to be miss-classified and hence these are over sampled to increase their prediciton accuracy.
Safe Level SMOTE	It over samples minority instances along the interpolated line between minority samples with different weight degree, called safe level. It remedies an issue with SMOTE ignoring nearby majority class samples. The Safe levels are computed using neighbouring minority and majority samples.

TABLE I
DESCRIPTION OF TESTED OVERSAMPLING METHODS FOR OUR BASELINE MODEL

and Safe Level SMOTE. As such ADASYN was used as the sampling approach in our baseline results.

Our improvement to the sampling process implemented by the seed paper was to treat the dataset as a timeseries dataset. A time series is a dataset where the data points are indexed in time order, ie; have a temporal ordering in the data space. This is important as successive data points in a time series are usually highly correlated, ie; have an effect on each other. Handling this correlation while sampling can help generate synthetic samples that are more representative of our dataset and it's corresponding temporal order. As noted in the feature generation section, we have two categorical features that are highly important for model prediction. The ADASYN approach cannot handle over sampling of categorical features.

To this effect we have chosen two methods that will handle the temporal order of the dataset and categorical features when sampling.

*1) **OSTSC**:* OSTSC is a R implementation by Dixon et al. [9] of Cao et al. [8] paper "Integrated Oversampling for Imbalanced Time Series Classification". We have detailed out a general outline of this method in our literature review section. It has two major improvements over the SMOTE variant sampling approaches on a timeseries dataset:

- It uses the enhanced structure preserving method to preserve the distribution structure of the minority samples in the data space and also uses a spectral filter to reduce noise when oversampling.
- Secondly the parents chosen for generating the new synthetic samples are highly temporally correlated as can be seen in Figure 8. The parents are chosen only in a given nearest neighbour region (shaded circle). This ensures that the generated synthetic samples are temporally correlated with the parent samples.

An important issue with this sampling approach is that it does not handle over-sampling for categorical features well. A robust solution would be to extend the categorical feature sampling process from SMOTE-NC to the OSTSC pipeline.

As a more quicker solution, we applied the OSTSC sampling method on the categorical features and then binned the generated values in ranges [0,2] for *feature: on_offshore* and [0,1] for *feature: BeachType*.

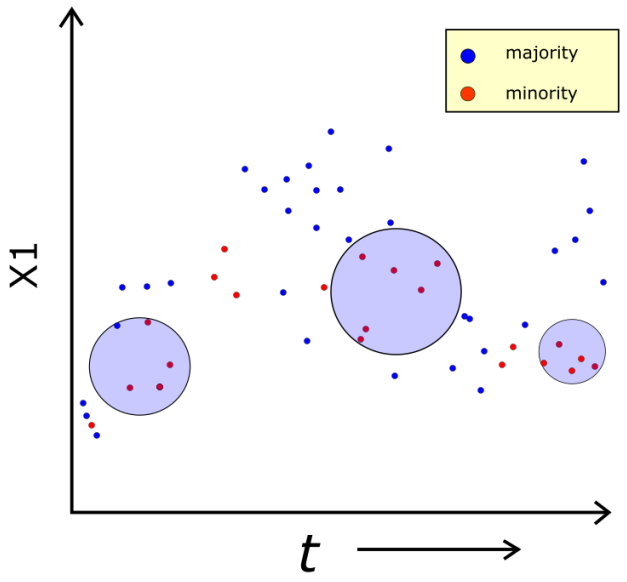


Fig. 8. OSTSC sampling

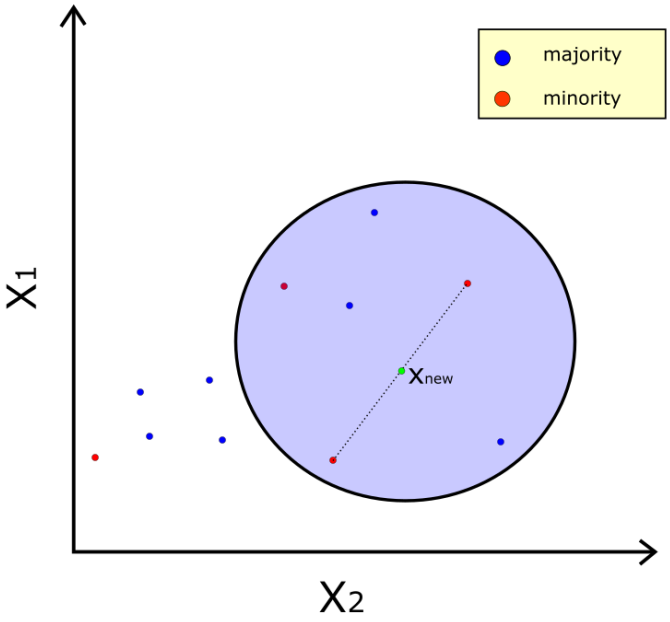


Fig. 9. SMOTE-NC sampling

*2) **SMOTE-NC**:* SMOTE-NC is an extension to the original SMOTE algorithm and addresses the over-sampling for categorical features. The categorical features of the new generated samples are decided by picking the corresponding most frequent category of the nearest neighbors present during the generation.

An issue with this sampling approach is that it does not take into account the temporal order of the samples. The parents chosen for synthetic sample generation can be from different timestamps in the data space 9. To address this issue, we applied sampling method in each training time split during the cross-validation process so that the parents are usually in the same temporal region. We have decided to use SMOTE-NC for the sampling process. This will be detailed in the next section.

D. Time Series Experimental Design

The data was he data was first sorted in date order and the first 90% of the data was used to train the model with the last 10% was used to test the model. The training data was then normalisation and the scalers used in the training normalisation were also applied when normalising the test data. This was done to ensure that no information from the test set managed to leak into to our training data. The normalised data was then used for training, validation and testing. Pythons SciKit Learn SMOTENC package was used for over-sampling the training data as this implementation can deal with both numerical and categorical features.

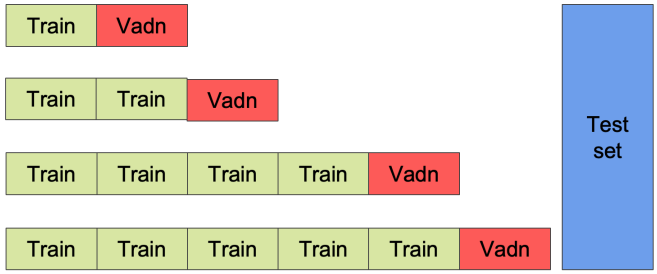


Fig. 10. Multiple Time Split

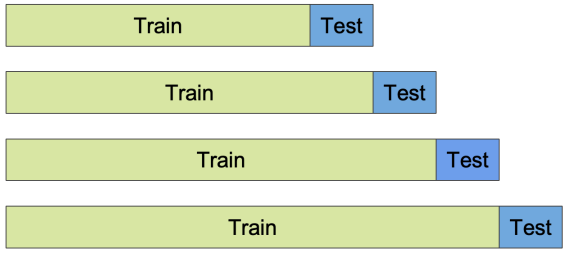


Fig. 11. Walking Forward

An important part of training a model is the tuning of hyperparamaters and each model has its own specific paramaters that need tuning:

- For KNN, as the class label is based on the class label of the majority votes from the neighbourhood, the number of neighbours in a neighbourhood is a crucial parameter.
- For BDT, unlike the other members from the tree family, it can overfit when we grow too many trees in the forest so the tree depth is a tuned parameter

- For ANN, the learning rate is the most important hyperparameter to rule it all. However, as our dataset was very limited we found we did not have enough data to successfully tune this paramater. Therefore, we used the sci kit learn default adaptive learning rates which converge faster without missing the optimal points or getting stuck in any local minimums points in the points inhigh-dimension model space

In order to retain any time signal information the hyperparameters were trained using cross validation techniques that maintain the time series order. Cross validation is a statistical resampling technique designed in such a way that each equally and randomly partitioned data (test sets) is evaluated against the remainder of data (train sets) to perform analysis. As we have a data set spanning across 20 years worth of water quality samples the experiments were designed so that the validation/test set always came after, in time, of the training sets. In this way, the temporal information could be reserved as much as possible with the focus on the prediction of the most recent data ("future" data for from the validation/test sets). To that end, we introduce multiple time splits for the train/validation sets and walking forward for train/test sets.

In the train phase multiple time splits 10 respect temporal order of observations by splitting the training set into equal-sized chunks chronologically. In the first cross validation chunk, the first chunk is used to evaluate against the second chunk. In the next round, the first two chunks are used to evaluate against the third one and so on until the last chunk is used as the validation set. When tuning the hyperparameters we chose to focus on sensitity as that is normally the hardest metric to obtain with unbalanced data. Also from a the health perspective, false negatives, that is predicting that water quality is safe but it's actually very dangerous, is the most important metric.

Once hyperparamaters were trained we used a walking forward 11 to evaluate the model on the test data. In this process the test set is split into equal-sized chunks chronologically, the first chunk is tested and then this chunk is incorporated into the training data and the next chunk is tested. In this way, the model is trained on the most recent data before testing. This is appropriate given our initial analysis indicated that there was and increasing trend in Entero levels. Figure 10 show the experimental design for both cross validation and testing.

IV. Results

A. Baseline Models: Replication of Seed paper models

As part of our replication of the seed paper, We applied 4 models that the seed paper chose (KNN, SVM, Boosted Decision Tree (XGBoost) and ANN) on the all beach dataset and for each of the individual 5 beaches. The ADASYN sampling approach was chosen to over sample the dataset. The original 7 features were used (Rain24, Rain48, Rain72, Wind Direction, Wind Speed, Solar Hours) and **the dataset is not considered a time-series**. It is a binary classification and the task at hand is to classify a given sample as below-threshold or above-threshold.

The KNN model with $N = 5$ showed the best results. For the SVM model we used the grid search process from Hsu et al [15] to find the best hyperparamters. We used a radial basis kernel as the features are not linearly separable. The grid search with cross-validation was applied on the parameters Cost $C = [2^{-}5, 2^{1}0]$ and Gamma $\gamma = [2^{-}5, 2^{0}]$. We got $C = 1$ and $\gamma = 0.3289722$ as the best parameters when modelled on the all beach dataset. For the XGBoost model, we got the hyperparameters, $eta(learningrate) = 0.001, max\_tree\_depth = 15$ and 5-fold cross validation method was used. The ANN was trained with a learning rate of 0.015 and no. of hidden neurons in the hidden layers = 8. Results are show in Table II.

Overall we were able to replicate the seed paper's results for KNN, SVM and ANN, with KNN and ANN performing the best in general as noted in the seed paper. We achieved minor improvements for the SVM model when compared to the seed paper. The Boosting Decision Tree results were comparatively much worse, though this could be a case of using XGBoost rather than GentleBoost and not sufficient hyperparameter tuning done during the cross-validation phase.

B. Baseline Model: multiple logistic regression

As a baseline general model for all beach sites we fitted a multiple logistic regression (MLR) with spline smoothing. Splines were chosen based on an inspection of the scatterplot of each variable against Enterro Class. The features used were those indicated from the feature selection algorithm. Based on the scatterplot a smoothing spline of 2 was used for the rainfall variables and the remaining variables were not smoothed. The first regression model used the actual unbalanced data set. As MLR with smoothing splines does not require transformed data we used the actual non-standardised data.

The first MLR model achieved an accuracy of model accuracy of 85%, model Sensitivity of 6.25% and model specificity of 100%. The model sensitivity was very low due to the unbalanced data set. The second MLR model was trained on data that was oversampled using R's SMOTE package. The package also standardises the data prior to generating synthetic samples but transforms the sampled back to the original value scales. The SMOTE generated data was balanced with 1,698 negative classes and 1,638 positive classes. The MLR model trained on the oversampled data achieved an accuracy of 75%, Sensitivity of 53% and specificity of 79%. This clearly demonstrated the increase in sensitivity due to balancing the training data, although this was at the expense of accuracy which reduced by 10% and specificity which decreased to 79%.

Accuracy Sensitivity Specificity MLR Original Data 85.15% 6.25% 100% MLR SMOTE Balanced Data 74.75% 53.13% 78.82%

C. Improvement: Machine Learning Models

Three machine learning models, K nearest neighbours, Boosting Decision Tree and Artificial Neural Network were trained on data oversampled using Python's SciKit Learn

Model	Clarks	Judges-bay	Narrow-neck	Weymouth	Mildford
KNN	**0.923, 0.741, 0.917**	0.864, 0.807, 0.944	**0.893, 0.824, 1.000**	**0.726, 0.72, 0.732**	0.889, 0.838, 0.956
SVM	0.884, 0.846, 0.923	**0.863, 0.909, 0.819**	0.785, 0.857, 0.714	0.692, 0.630, 0.753	**0.907, 0.851, 0.962**
XGBoost	0.884, 0.770, 1.000	0.820, 0.682, 0.955	0.750,0.571, 0.929	**0.719, 0.644, 0.795**	0.870, 0.778, 0.963
ANN	- , 0.833, 0.944	- , 0.802, 0.921	**- , 0.855, 0.960**	**- , 0.760, 0.785**	**- , 0.844, 0.920**

TABLE II
BASELINE RESULTS. THE MODEL SCORES ARE GIVEN IN THE FORMAT ACCURACY, SENSITIVITY (TRUE POSITIVE RATE), SPECIFICITY (TRUE NEGATIVE RATE)

Model setup	Accuracy	Sensitivity	Specificity
MLR Original Data	85.15%	6.25%	100%
MLR SMOTE Balanced Data	74.75%	53.13%	78.82%

TABLE III
MLR RESULTS

SMOTE package. Feature included were those selected by the Feature selection algorithm. The table below shows the mean metrics for each model averaged over the four test splits. KNN had the highest mean accuracy and sensitivity and ANN had the highest mean sensitivity. In order to see if these differences were significantly different we ran SciKit Learns autorank on the results. For all three measures there was no evidence of non-normality or unequal variance in the results scores so a ANOVA test for any difference was conducted along with post hoc tukey tests. According to autorank output, ANN significantly lower than KNN BDT for accuracy. BDT was significantly lower than ANN KNN for sensitivity and for specificity ANN was significantly lower than KNN BDT. On that basis we concluded that the KNN model was the best overall. However it should be noted that with only 4 data points to compare for each model the autorank method is probably unnecessary.

V. DISCUSSION

While our final trained models fell short of SafeSwim's required 85% accuracy and 80% specificity, with our best average of 75% and 72% respectively, our models were able to achieve a much higher degree of sensitivity, with ANN model achieving a mean sensitivity of 80% and our best overall model, KNN, achieved 77% sensitivity. These values are in excess of the 50% sensitivity required by safe swim and we deliberately aimed to get a high sensitivity above all other metrics. We aimed for high sensitivity as this has been the trickiest metric to improve and it is the most important in terms of an overall health objective which is the main objective of the safeswim campaign. The accuracy and sensitivity can be increased by reducing the amount of oversampling in the training set but this is at the expense of sensitivity.

A. Challenges

A big challenge in this research was limitations in the sampling data. Firstly there was only 2017 data points spread over 20 years and within that time the sampling methods changed and the population density in Auckland has increased significantly placing a higher burden on stormwater and sewage infrastructure. In addition we discussed sampling methods with the laboratory technician involved in sampling the data and she suggested that there were big differences in both the number and nature of sampling during the data period. We feel that this probably had a detrimental effect on our modelling results. A 2018 study [safeswim] assessing the success of the safe swim campaign found that there was serious issues with the sampling data. The authors found that there was an inherent bias in the sampling data and that samples were much more likely to be taken during fine weather conditions due to health and safety concerns. In addition water samples from one beach (not included in our study) only found one incidence of dangerous Entero levels during the sampling period whereas targeted sampling at a nearby stream at the same time found elevated, unsafe levels. The authors suggested that the sampling regime prior to 2017 was probably understating the health risk at that beach. In light of this research the sampling methodology was overhauled and a new sampling regime was put in place.

We did try to include more recent data in our study and managed to source sampling from 2016 to 2020 however when we tried to integrate this data into our sampling data set we found differences in the rainfall amounts from the two different data sets even though these observations were taken on the same day.

B. Future Work

As part of this research we talked to the environmental department of Auckland Council, the department responsible for the safeswim water sampling and beach notifications. They indicated that the tide height at the time of any bacteria contamination was had a big influence on bacteria levels. That is if the tide was going out at the time of contamination then bacterial levels were likely to be lower as the bacteria was carried out with the tide. The reverse is true if the tide was coming in. However the sampling data that we had did not record the time that the sample was taken so we were unable to assess this. We note that the new sampling regime includes a time when the sample was taken and that samples were taken around the high tide level in order to standardise this. We feel that this would be a key feature to include in further research. In addition in order to build a general model we feel that 5 beaches were probably insufficient. Particularly given the fact that each of these 5 beaches had some characteristic that was different from all others. Weymouth was close to a sewerage treatment plant, the east coast beaches on the north shore were both coastal whereas the southern beaches were

more tidal. More beach sites could have identified predictive features more easily.

VI. Conlusion

We could not have modelled the data as well without talking to domain experts. We initially made some incorrect assumptions around sampling techniques and data quality and we did not realise this until we talked to Auckland Council and a laboratory technician.

In conclusion we feel that it is possible to build a general model for all Auckland beaches that would meet the safeswim's metric requirements and we were able to acheive a high level of sensitivity using oversampling techniques and controlling for time signals in the data. However Auckland's stormwater infrastructure is a key factor in determining bacteria contamination as outlets with older systems in high density areas regualry overflow. Therefore some specific beach characteristics such as location should alway be included.

References

[1] D. S. Fancy, *Developing and implementing predictive models for estimating recreational water quality at Great Lakes beaches*. US Department of the Interior, US Geological Survey, 2013.

[2] T. Xu, G. Coco, and M. Neale, "A predictive model of recreational water quality based on adaptive synthetic sampling algorithms and machine learning," *Water research*, vol. 177, p. 115788, 2020.

[3] Z. Zhang, Z. Deng, and K. A. Rusch, "Modeling fecal coliform bacteria levels at gulf coast beaches," *Water Quality, Exposure and Health*, vol. 7, no. 3, pp. 255–263, 2015.

[4] C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano, "Mining data with rare events: a case study," in *19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007)*, vol. 2. IEEE, 2007, pp. 132–139.

[5] B. Krawczyk, "Learning from imbalanced data: open challenges and future directions," *Progress in Artificial Intelligence*, vol. 5, no. 4, pp. 221–232, 2016.

[6] J. Kong, T. Rios, W. Kowalczyk, S. Menzel, and T. Bäck, "On the performance of oversampling techniques for class imbalance problems," *Advances in Knowledge Discovery and Data Mining*, vol. 12085, p. 84, 2020.

[7] J. Alcalá-Fdez, A. Fernández, J. Luengo, J. Derrac, S. García, L. Sánchez, and F. Herrera, "Keel data-mining software tool: data set repository, integration of algorithms and experimental analysis framework." *Journal of Multiple-Valued Logic & Soft Computing*, vol. 17, 2011.

[8] H. Cao, X.-L. Li, D. Y.-K. Woon, and S.-K. Ng, "Integrated oversampling for imbalanced time series classification," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 12, pp. 2809–2822, 2013.

[9] M. F. Dixon, D. Klabjan, and L. Wei, "Ostsc: Over sampling for time series classification in r," *Available at SSRN 3077767*, 2017.

[10] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *The Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.

[11] U. of Minnesota, "Table 3: Wind direction and degrees," Available at http://snowfence.umn.edu/Components/winddirectionanddegrees.htm.

[12] N. W. Service, "Origin of wind," Available at https://www.weather.gov/jetstream/wind.

[13] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of machine learning research*, vol. 3, no. Mar, pp. 1157–1182, 2003.

[14] D. Camp, "Beginner's guide to feature selection in python," Available at https://www.datacamp.com/community/tutorials/feature-selection-python.

[15] C.-W. Hsu, C.-C. Chang, C.-J. Lin *et al.*, "A practical guide to support vector classification," 2003.