FEATURE SELECTION AND FUZZY CLASSIFICATION FOR THE PREDICTION OF CO2 RATINGS

Lappeenranta–Lahti University of Technology LUT

Master's program in Business Analytics, Engineering Science, Master's Thesis

2023

Shageerthana.Sathiyamoorthy

Examiners: Professor Pasi Luukka

Assistant Professor Tapio Helin

ABSTRACT

Lappeenranta – Lahti University of Technology LUT

School of Engineering Science

Computational Engineering - Business Analytics

Shageerthana Sathiyamoorthy

Feature selection and Fuzzy classification for the prediction of CO2 ratings

Master's Thesis

2023

49 pages, 15 figures, 8 tables

Examiners: Professor Pasi Luukka

Assistant Professor Tapio Helin

Keywords: Feature selection, Fuzzy entropy, Fuzzy kNN, ANFIS, Similarity based classification, Confusion matrix, Performance measures, Multi-class classification.

The rapid growth of technology has led to the generation of large amounts of data in all possible sectors. There are varieties of algorithms that can be used to analyze and learn about the data in the machine-learning world. It is a best practice to preprocess the data before it undergoes any kind of learning. Feature selection is one of the steps done in the data preprocessing. In this study, we will be focusing on the fuzzy entropy-based feature selection method together with 3 different fuzzy classifiers. The features selected from the dataset will be tested in FkNN, Similarity-based, and ANFIS classification models. The accuracies were obtained from each model. Based on the highest accuracies FkNN 98%, Similarity 95%, and ANFIS 97% obtained, respectively. The performance measures such as sensitivity, specificity, and precision values were examined for each model and it was found that the classes were imbalanced and all 3 models failed to predict the positive incidents, while the models predicted the negative incidents more accurately.

ACKNOWLEDGEMENT

I am thankful to my family who believed in me and keep on encouraging me until this point of my master's study. Also, my friends who have been there whenever I needed in all possible ways and did not let me to lose hope in myself.

I would like to thank my supervisor Prof. Pasi Luukka for the guidance and support that he gave throughout my thesis. I should say that without his motivation and suggestions on how to carry out this thesis I would have not been able to complete it promptly.

Finally, I appreciate Lappeenranta-Lahti University of Technology (LUT) for offering me the opportunity to do my master's degree with a scholarship. Throughout this journey I learned a lot in my field of studies, and I believe I will be able to make use of these in my career.

Lappeenranta, August 6th, 2023

Shageerthana Sathiyamoorthy.

ABBREVIATIONS

kNN	k nearest neighbors
ANFIS	Neuro Fuzzy Inference System
FkNN	Fuzzy k nearest neighbors
FIS	Fuzzy Inference System
RFE	Recursive feature elimination
LOOCV	Leave one out cross validation
SVM	Support Vector Machine
MI	Mutual Information
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
CO2	Carbon dioxide

Table of Contents

1 Introduction 7
1.1 Objective 8
1.2 Structure of the thesis 9
2 Literature Review 10
2.1 Chronicles of feature selection 10
3 Theoretical understanding of mathematical model 12
3.1 Crisp set and fuzzy set 12
3.2 Fuzzy measures 13
3.3 Fuzzy entropy measure 13
3.4 Feature selection 14
3.4.1 Filter method 14
3.4.2 Wrapper method 16
3.4.3 Embedded method 17
3.5 Other studies related to feature selection 19
3.6 Classifiers 20
3.6.1 FkNN classifier 20
3.6.2 Similarity based classifier 22
3.6.3 ANFIS classifier 23
4 Methodology 26
4.1 Data Acquisition 26
4.2 Descriptive statistics 27
4.3 Data Preprocessing 28
4.4 Data Sampling 29
4.5 Performance Measures 30
5 Results 32
5.1 Data 32
5.2 Results and findings 32
5.2.1 Results from FkNN model 34
5.2.2 Results from the Similarity classification 35

5.2.3 Results from the ANFIS classification 37
6 Discussion 39
7 Conclusion 40
References 41

1 Introduction

The contemporary advancement in technology has led to the growth of data driven world. Where we have immersed ourselves into it and have become inevitable to ignore the presence of world revolutionized and interconnected by data. In this context there are various studies being done on collecting the essential data (Sarker, 2021). Data mining is considered to be one of the highly used methods in the context of machine learning and artificial intelligence. It can be divided into Supervised learning and unsupervised learning methods. Unlike the supervised learning unsupervised learning learn the data only from the inputs not from the outputs of the data; while supervised learning learn the data from both input and corresponding output of the data.

High dimensional data has become as one of the topics in many sectors due to the rise of automated data collection techniques. High dimensional data is widely found in the health sector (immune system status, blood pressure, height, weight etc.), in financial data and so on. Because of the complexity of the high dimensional data the requirement for more memory space and high computational cost and multicollinearity have become a problem (Aggarwal et al., 2001). As a solution for this problem dimensionality reduction can be done. Feature extraction, feature selection is some of the remedies used for these conditions (Li et al., 2018). By reducing the number of features available in the dataset, feature extraction seeks to produce new features. Using these extracted features, the original dataset will be summarized. On the other hand, feature selection is where a subset will be selected from the original feature set.

It is one of the vital step to conduct feature selection in machine learning as it helps to reduce the computational time and to obtain good predictions (Chandrashekar & Sahin, 2014). Feature selection does not always result in improving the accuracy, but it is recommended to implement this step in every machine learning problem which will be helping in some ways, it can be either in reducing the computational time, reducing the complexity of the datasets, shrinking the memory requirements etc. It has been stated that the correct feature subsets have the ability to minimize the measurement quantity, the performance duration will be reduced resulting in improving the clarity of the model (Luukka, 2011). Feature selection helps in neglecting the noises or irrelevant datapoints found in the datasets in some cases it can also help to prevent overfitting of the model.

There are several classification models used to classify and analyze different data. The development of fuzzy kNN model (Keller et al., 1985) , Similarity based classification (Luukka et al., 2001) and Adaptive neuro fuzzy inference system (ANFIS) (Jang, 1993) are some of the useful methodologies. These mentioned fuzzy classification models are used in this study in the MATLAB environment in which the fuzzy entropy feature selection method is used.

1.1 Objective

The main intention of this study is to perform the feature selection using the fuzzy entropy method used in (Luukka, 2011) by testing and comparing its compatibility in 3 different fuzzy models namely, FkNN, Similarity based classification and ANFIS.

Delimitations

- The research is delimited to a certain multi class dataset analysis. Hence the predictions and observations obtained are not generalize to other datasets having different data types, characteristics, sizes, or domains.
- The feature selection is demarcated with the usage of fuzzy entropy method with FkNN, similarity-based classification and ANFIS modeling. But there are several other techniques for feature selection which are not included in this study.

Limitations

- The classification models are trained and tested on a specific multi class dataset. The results will not be application for other models or other types of multi class datasets.

The study will cover the following questions,

1. How does the incremental removal of the features using fuzzy entropy impact the performance of each model?
2. How do different machine learning models behave when same task is applied?

1.2 Structure of the thesis

The section briefly explains how the thesis report has been organized.

Chapter 2 Contributes to the literature review of the studies and experiments done so far in the focus topic

Chapter 3 Dives into the mathematical background of the research. With the provision of mathematical formulas.

Chapter 4 This consists of the implementation of the data pre-processing, feature selection and illustrating the performance of the models.

Chapter 5 Speaks about the derived findings from this research.

Chapter 6 Illustrates the current study.

Chapter 7 Summarizes the research work and possibilities of future work.

2 Literature Review

Classification is known as one of the fundamental study areas in machine learning and data mining (Baradwaj & Pal, 2012). Classification models predict the future outcomes based on a set of known inputs (Muhammad & Yan, 2015). With the help of a mapping function the prediction of the class label is done corresponding to the test data from the training dataset. This will become the bridge for the predictive model, due to the feature being related to the dataset.

The classification and predictive models have already been proving their contributions and capabilities in the healthcare research and applications (Hassani et al., 2017). These are considered as an important tool in the disease diagnosis and prognosis helping to find the patterns of the diseases. For example the predictive models have been used to analyze the diabetes patients' data and predict it (Fernandez De Canete et al., 2012) etc.

Due to the rapid growth of large number of data the feature selection paves a key role in the machine learning (Guyon et al., 2003). It was stated by (Kira & Rendell, 1992) that the feature selection technique has been implemented often to reduce the dimensionality of the datasets. Due to the presence of the curse of dimensionality the traditional learning algorithms fail to scale the large datasets. Noisy features always tend to produce vague results in the model affecting the performance of the models negatively, this problem is mitigated with the help of feature selection (Huan Liu & Lei Yu, 2005).

2.1 Chronicles of feature selection

Feature selection has become one of the highly researched topic both in practice and in methodology for many years (Blum & Langley, 1997), (Kira & Rendell, 1992). The removal of the feature will never be a drawback to the performance of the any machine learning models (Li et al., 2018). In the studies of (Dash & Liu, 1997) they have highlighted the need for dimensionality reduction and the performance for classification algorithms. Providing intuitive information about various techniques related to feature selection and presented their potency with experimental results.

Concept learning algorithms have had an immense impact on the machine learning world, paving a path in feature selection techniques. The main goal of this algorithm is to describe the concept using feature selection (Mitchell & McGraw, 1997). Following are some of the related studies conducted back in the days,

- Decision trees : (Quinlan, 1986) ID3, (Quinlan, 1993) C4.5 proposed effective methods for the feature selection using the information gain and gain ratio. The most important features found at the top of the tree are the features used to separate the nodes. The study of gain ratio and gain information were used by (Rendell & Seshu, 1990) states that there will be a conflict between the features that will arose if induction algorithms are introduced in the concept learning algorithms. Here the induction algorithm means: A number of training data in which each and every data observation is depicted by a vector of feature values and class labels.
- Find S – Algorithm : This learning is based on conjunctive concepts (Mitchell & McGraw, 1997) But it does not directly conduct the feature selection indirectly delivers the idea through generalizing the necessity of the concept by hypotheses.
- Perception algorithm : This is considered to be one of the primitive algorithms (Rosenblatt, 1958) mainly used for linearly separable data sets.
- Candidate elimination algorithm : Yet another study related to conjunctive concepts (Mitchell & McGraw, 1997) by maintaining the most common and specific hypothesis bringing up the accurate concept.

These methodologies can be considered as the earliest studies comparatively to those techniques used nowadays, obviously being the foundation for the feature selection process.

3 Theoretical understanding of mathematical model

Large number of mathematical models were implemented over the years of scientific studies in order to simplify the complex processes and to provide a better understanding. This section will give the mathematical idea of the fuzzy sets not leaving the fuzzy entropy which has been used in the studies for the features selection.

3.1 Crisp set and fuzzy set

If *A* is a crisp set defined over a universe *X*, according to the set theory the element found in the set is either a member of *A* or not. This theorem is explained using membership function $\mu_A(x)$ with two values; 1 representing that if $x \in X$ is a member of *A* and 0 vice versa.

$$\mu_A(x) = \begin{cases} 1, & x \in X \\ 0, & x \notin X \end{cases} \quad (1)$$

Fuzzy set is a derived and generalized version of the classical crisp set. In which the above-mentioned property is generalized by accepting partial membership of a set. Fuzzy logic is a derivation from the fuzzy set theory. Hence the fuzzy set *A* is related to the membership function $\mu_A(x)$, includes all the elements found in the universe of discourse *X* in the interval of [0,1]. In other words, we can say that fuzzy sets consist of elements that would persuade imprecise properties of membership. The definition of fuzzy set can be,

$$\mathrm{A} = \{(x, \mu_{\bar{A}}(x)),\ x \in X\} \quad (2)$$

Fuzzy set *A* which is defined on *X* is presented as in the equation if *X* is a universe of discourse and *x* will be the element of *X*.

Fuzzy set is defined by the membership function. A measure of degree of similarity of an element is provided by the membership function to the fuzzy set. Accordingly, there are different membership function varying with the shapes.

- Triangular membership

  Described using 3 parameters {a, b, c}, where these represent the x coordinates of $\mu_A(x)$ vertices in a fuzzy set.

$$\mu_A(x) = \begin{cases} 0 & if\ x \leq a \\ \frac{x - a}{b - a} & if\ a \leq x \leq b \\ \frac{c - x}{c - b} & if\ b \leq x \leq c \\ 0 & if\ x \geq c \end{cases} \quad (3)$$

3.2 Fuzzy measures

C is the class of subsets found in universe discourse *X*. *G*, *F* ∈ *C* so *g*: *C* ⟹ ℝ, this is known as fuzzy measure.

$$\emptyset \in C \Rightarrow g(\emptyset) = 0 \quad (4)$$

$$G \subseteq F \implies g(G) \leq g(F) \quad (5)$$

3.3 Fuzzy entropy measure

This is introduced by (De Luca & Termini, 1972) and can be represented as follows, where number of elements is *n* in *A* and *L* is a constant.

$$H(A) = -L \sum_{i=0}^{n} \mu_A(x_i) \ln \mu_A(x_i) \quad (6)$$

In a similar way they brought an equation known as entropy of fuzzy set

$$F(A) = H(A) + H(A^c) \tag{7}$$

With a log function the fuzzy entropy H was implemented, and the equation goes as follows,

$$H = -\sum_{i=0}^{n} (\mu_A(x_i) \ln(\mu_A(x_i)) + (1 - (\mu_A(x_i)) \ln(1 - \mu_A(x_i)) \tag{8}$$

3.4 Feature selection

The definition for feature selection given by (Bishop, 2006) it is the list of numbers of the columns of tables. According to (Huan Liu & Lei Yu, 2005) feature selection is the process of selecting a subset of original features from a dataset. Feature selection method can be classified into three divisions namely : Supervised methods (Wolf & Shashua, 2003), Semi supervised methods (Zhao & Liu, 2007) and Unsupervised methods (Cai et al., 2010). Supervised feature selection can select the relevant features with the availability of labelled information. While the semi supervised feature can work properly with limited number of features in terms of labelled and unlabeled data. Unsupervised feature selection is typically not easy to handle due to the absence of labels.

The feature selection is categorized into 3 main divisions such as Filter method, Wrapper method and Embedded method based on the methodology. These methods are usually employed under the supervised feature selection technique (i.e., it focusses on the target variables and the labelled dataset it uses for this purpose).

3.4.1 Filter method

Features are selected based on the statistical measures and in the preprocessing stage. The unwanted features / the features which are not redundant for the model will be removed by means of various metrics (Blum & Langley, 1997). There are no feature dependencies available in this

method due to the tendency of the interaction with the classifier. But this method comparatively requires low computational time and as an added advantage it eliminates the overfitting of the model.

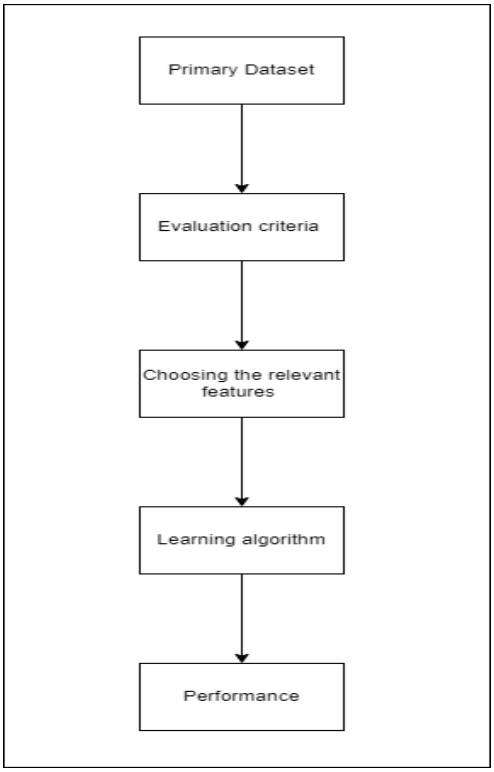

Figure 1 Filter method

The figure 1 shows how the filter method functions. The statistical measures are calculated in order to determine the relevance of each, and every feature inputted in the initial stage of the process. Subset will be chosen based on the feature ranking, and these selected features will be fed into to model.

The features can be determined either by univariate or multivariate way. As per the terminology for univariate method the features are ranked singularly and in multivariate method several features are ranked (Mehnaz Khan & S. M. K. Quadri, 2013).

The filter method was introduced initially by (Hussein & Thomas G, 1991) and (Kira & Rendell, 1992). In which the FOCUS algorithm was used that could learn the concepts of a variety of irrelevant features. It is based on the Occam's razor principle which decides the best data with the

simplest hypothesis. The combination of the relevant features and creating a hypothesis is the first step. The application starts by learning simple hypotheses implying a single feature followed by multiple features creating a complex hypothesis. A divide and conquer strategy have been used to perform all these steps. Through their experiments they concluded that this FOCUS algorithm outperforms the traditional ID3 (decision tree) as it identifies the target concepts even in the absence of perfect features. In other terms the curse of dimensionality was addressed through this study.

Another implementation of (Kira & Rendell, 1992) was the RELIEF algorithm for the feature filtering. This works on a single instance rather than on a highly abstract dataset. The quality of the features is calculated based on the values that are closer to each other. This is done with the combination of the instance-based algorithm. They demonstrated if a feature is owning a high weight, it is highly relevant while its low weight it is not applicable for the classification. Through their numerous experiments that RELIEF worked well with features with noises and when they are relied on each other.

3.4.2 Wrapper method

Wrapper method considers the feature selection as a searching issue, various combinations are created and evaluated. The algorithm is trained using the subset of features (John et al., 1994). Features are selected and eliminated based on the output of the model. Following are the techniques of wrapper method.

- Forward selection : This method initiates with empty features in which the model's performance will be improved with the inclusion of a number or features. (Saeys et al., 2007).
- Backward selection : This method removes the irrelevant feature one after the other where at the initial stage all the features will be used. (Guyon et al., 2002)
- Exhausted feature selection : This method performs in a combination manner where the best combination of the features is selected. (Somol et al., 2006)
- Recursive feature elimination : RFE is a special method which is similar to the backward selection method used to remove the unnecessary features one by one (Guyon et al., 2002)

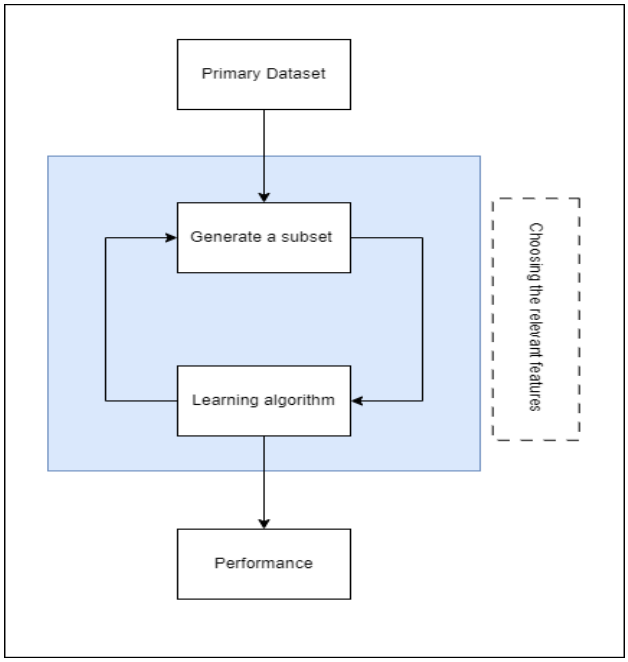

Figure 2 Wrapper method

As you can see in the figure 2 The subsets will be generated once the dataset is obtained unlike the filter method. Whereas these subsets are then used to build a model followed by the performance evaluation. The features are selected based on the subsets which produced the best performance model, and those features are then used for further model development.

According to the studies done by (John et al., 1994) on induction algorithm where the feature subset was applied as a wrapper, asserted that wrapper method performs way better than filter metho. Because of the collaboration with the classifier directly the wrapper method is considerably better than filter method with providing good accuracies. The presence of induction algorithm let it to provide good performance (Kohavi & John, 1997). Even though it has good qualities, there are some bad qualities as well. Overfitting is one of the biggest disadvantages of this method and computational cost is high this makes several problems out of reach.

3.4.3 Embedded method

During the modelling algorithm's execution stage, the feature selection is performed by the Embedded method. The methods are embedded as normal or as an extended functionality in the

algorithm. The most common methods are CART, C4.5, random forest coming under the types of decision tree algorithm (Zani et al., 2006). The most informative and relevant features are identified as embedded method is a combination of filter and wrapper method (Li et al., 2018).

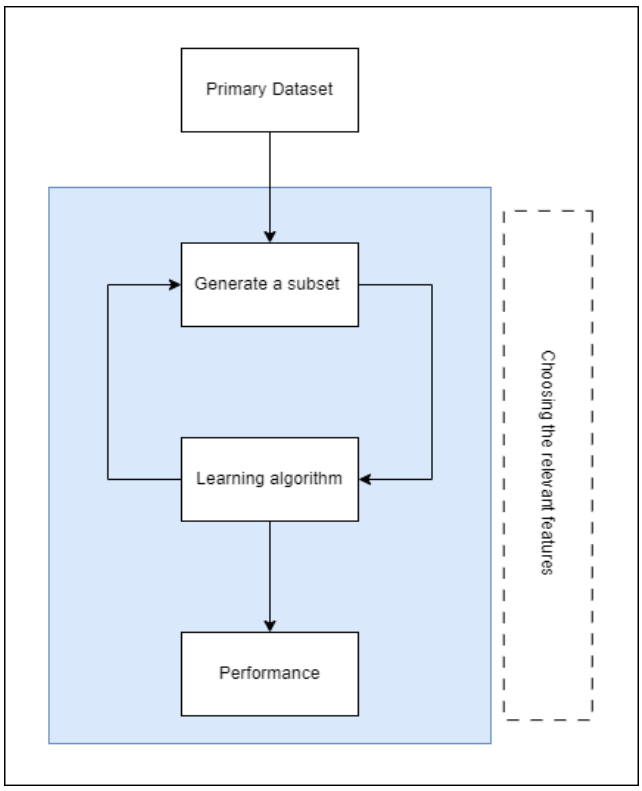

Figure 3 Embedded method

As you can see in Figure 3 the model construction method for the embedded method starts as soon as the data is fed into the model. While this process is undertaking, the learning algorithm will perform the evaluation of the features as a part of the model building process. Finally, the features are selected based on the highest model performance and used in the model building.

This method possess several advantages such as comparatively lower computational power (Blum & Langley, 1997), it is not necessary to split the data into training and testing data (Guyon et al., 2002) which is why it is considered an efficient method. The major disadvantage of this method is the dependency it has over the classifier even though it provides some good results (Mehnaz Khan & S. M. K. Quadri, 2013). An embedded feature selection method implemented by (Guyon et al.,

2002); involving the Support Vector Machine (SVM) together with Recursive feature elimination (RFE). Based on the ranking obtained the features were removed recursive usage of SVM -RFE.

3.5 Other studies related to feature selection

A study by (Jinjie Huang et al., 2006) used the mutual information (MI) in the local search operation to rank the features. SVM model is used for this purpose to achieve good results for the classification. i.e., MI is used to calculate how much of information essential to predict the correct feature on the target variable. SVM has been used also in the study of (Maldonado & Weber, 2009). Four different datasets were used and the features that will not impact the classification model were eliminated using the wrapper method. Precisely sequential backward elimination of features was done. A PRESET algorithm was used by (Modrzejewski, 1993) which is a type of filter method. In his experiment he used rough sets of heuristically ranked features, and those which features which were assumed as noiseless were eliminated.

A comprehensive ideology behind feature selection with the outcome of the selecting the best features have been mentioned in the research of (Li et al., 2018). He also mentioned that the feature selection algorithm can be used for four types of data such as flat features, steaming feature, heterogeneous features, and structure features. Studies of (Kohavi & John, 1997) made use of the statistical methods for feature selection by presenting it in the wrapper method. They showed that the feature selection is integrated with the learning algorithm will deliver a robust result with compared to the filter approach. (The feature selection will be handled as a preprocessing step in the data mining). A memory based wrapper method implemented by (Moore & Lee, 1994) was used to find the subset feature which was capable of highlighting the best feature. In particular as the subset evaluation function, leave one out cross validation (LOOCV) is used which also known as memory-based algorithm model.

The emergence of data dimensionality led to a new challenge in the feature selection. Where (Guyon et al., 2003) came up with various techniques that could handle this challenge dealing with text data or genomic data. Special characteristics of the bioinformatics data and the appropriateness of different methods of feature selection have been analyzed by (Saeys et al., 2007) in an extensive manner. An introduction to feature selection and variable pertinent to the machine learning

problems like regression and classification problems has been mentioned in the studies of (Guyon et al., 2003).

3.6 Classifiers

3.6.1 FkNN classifier

The fuzzy kNN theory was proposed by (Keller et al., 1985) which was derived from kNN classifier. The accentuates the FkNN classifier where it allocates a degree of membership to each class. Followed by predicting the class sample with regard to the highest membership degree (Keller et al., 1985).

Implementation of FkNN undergoes five different steps. Such as parameter initialization, calculation of the distance, determining the k nearest neighbors, computation of the membership and finally class prediction. It can be said that the first three steps are incorporated in a kNN algorithm. The fuzzy strength value is included as an additional parameter for FkNN classifier this is achieved in the first step with the initialization of the k value.

Let us consider a training set $\{ X_j , c_j \}_{j=1}^{N}$, In which $X_j = \{ x_j^1 , x_j^2, \ldots, x_j^D \} \in \mathbb{R}^D$ found in N instances of D – dimensional feature space of S classes. $c_j \in \{ \omega_1, \omega_2, \ldots, \omega_S \}$ here X_j is c_j . Using the kNN rule the class ω_c is set on in a query sample $Y \in \mathbb{R}^D$.

Step 1: Finding the number of *k* nearest neighbors. This is determined heuristically due to the nature of the algorithm. By making use of the training data different *k* values are tested for corresponding accuracies. And when the accuracy is optimal that particular *k* is chosen.

Step 2: Finding the neighbors of *Y* by measuring the similarity between *Y* and the entire training dataset. Usually this is done by using the Euclidean distances. It can be expressed as below,

$$d_{EUC} (Y, X_j) = \sqrt{(Y - X_j)^T (Y - X_j)} \quad (9)$$

Here the Euclidean distance between the datapoints *Y* and X_j is $d_{EUC} (Y, X_j)$ and j = 1, 2,…,N.

Step 3: The optimized k values with the class labels are used to identify the nearest training instances.

Step 4: in this step the fuzzy membership degrees based on the kNN rule are assigned to the initially for the nearest neighbors followed by the query instances. For every instance of X_j the fuzzy labelling the closest neighbors *K* will be found.

$$u_{ji}(X_j) = \begin{cases} 0.51 + (n_i/K) * 0.49, & if\ c_i = c_j \\ (n_i/K) * 0.49, & if\ c_i \neq c_j \end{cases} \quad (10)$$

The neighboring instances of X_j is indicated by n_i in the i^{th} class. After the generation of the class nearest neighbors, u_i(y) is calculated for the query sample $Y \in \mathbb{R}^D$.

$$u_i(\mathrm{y}) = \frac{\sum_{j=1}^{k} u_{ij} \left(1/\| y - x_j \|^{2/(m-1)}\right)}{\sum_{j=1}^{k} \left(1/\| y - x_j \|^{2/(m-1)}\right)} \quad (11)$$

Here in the i^{th} class of training data u_{ij} is the membership of the j^{th} sample and m $\in$ (1, +∞) where m = 2 always. $\| y - x_j \|$ is the Euclidean distance between the j^{th} neighbor of *X* and *Y*

Step 5: With regard to the highest membership degree received from all the classes the classification of *Y* into ω_c is determined.

$$\omega_c = \arg max_{\omega_c} u_i(Y), \quad i = 1,2,\ldots,S \quad (12)$$

There are some research done on feature selection based on FkNN. A bankruptcy prediction model was made using the feature selection together with FkNN classification by (Chen et al., 2011). As we know most of the time the classification models are made use in the medical fields. The model proposed by (Vommi & Battula, 2023) used the FkNN based on Bonferroni mean (Mailagaha Kumbure et al., 2020) was used as the learning algorithm in the feature selection using hybrid filter method such as Relief and fuzzy entropy (RFE) was developed for the COVID -19 case study.

FkNN offers several advantages when it comes to especially for uncertain / complex datasets. Unlike the kNN algorithm the FkNN has the ability to deal well with uncertain datasets. When it comes to multi class problems FkNN provide good performance compared to the traditional kNN.

3.6.2 Similarity based classifier

Similarity based classification was introduced by (Luukka et al., 2001). This classifier is based on fuzzy similarity measures. Łukasiewicz structure was used for defining the membership of the objects. It is said that Łukasiewicz structure is the only multivalued structure which have the characteristics of keeping the mean of many similarities as a similarity. The classifier calculates the class label of test sample with regard to the similarities between the set of labeled training samples and test samples. This classifier does not need to have a direct connection to the features found in the dataset. There is no strict rule to have a Euclidean space to be the sample space as long as the function is defined properly (Yihua et al., 2009).

If we consider how it is implemented : First the data is split into training and testing samples using the stratified sampling (Witten et al., 2011) with an 80:20 ratio. Making sure that the classes are distributed in both splits are similar to the whole dataset. The classification is done by classifying a set of X objects into N different classes: C_1, …,C_n by their features. Let us say D as the number of different kinds of features f_1,…., f_n which can be measured from objects. In order to have the features between a value of [0,1] it undergoes normalization. Which in turn says that the objects that are being classified are vectors of $[0,1]^D$. The ideal vector for each class i should be determined v_i = (v_i (1), …, v_i (D)), whereas this vector can be user defined or it can be calculated from some sample set X_i of vectors x = (x (1), …, x (D)) belonging to the class C_i. The v_i is calculated using the generalized mean, here for all i, d the power value of m is fixed.

$$v_i\,(\mathrm{d}) = \left(\frac{1}{\#X_i}\sum_{x\in X_i} x(d)^m\right)^{\frac{1}{m}},\ \forall \mathrm{d} = 1, \ldots, \mathrm{D} \tag{13}$$

After the determination of ideal vectors, the decision on which class an arbitrarily chosen x ∈ X should be is made by comparing it to each ideal vector. This comparison is done using the Łukasiewicz structure.

$$S\langle x, v\rangle = \left(\frac{1}{D}\sum_{d=1}^{D} w_d\,(1 - |\,x\,(d)^p - v\,(d)^p\,|\,)^{m/p}\right)^{1/m} \tag{14}$$

x, v ∈ [0,1]^d. p is the parameter present in the generalized similarity measure. If $S\langle x, v_m\rangle$ = max $S\langle x, v_i\rangle$ it is considered that x ∈ C_m.

Similarity based classification has been used widely in the computer vision, bioinformatics, processing of natural language, retrieval of information etc. According to (Santini & Jain, 1999) at some instances the similarity functions may be asymmetric and will blunder to work well with mathematical properties for metrics.

3.6.3 ANFIS classifier

Adaptive neuro fuzzy inference system (ANFIS) developed by (Jang, 1993) is based on Takagi-Sugeno model is a type of neural network. Fuzzy principles and artificial neural network (ANN) principles are used in Adaptive neuro fuzzy interface system (ANFIS). The choice of the number and the shape of the membership function determines the performance if the ANFIS classification (Talpur et al., 2017). The general the ANFIS model is represented by 2 rules-based system.

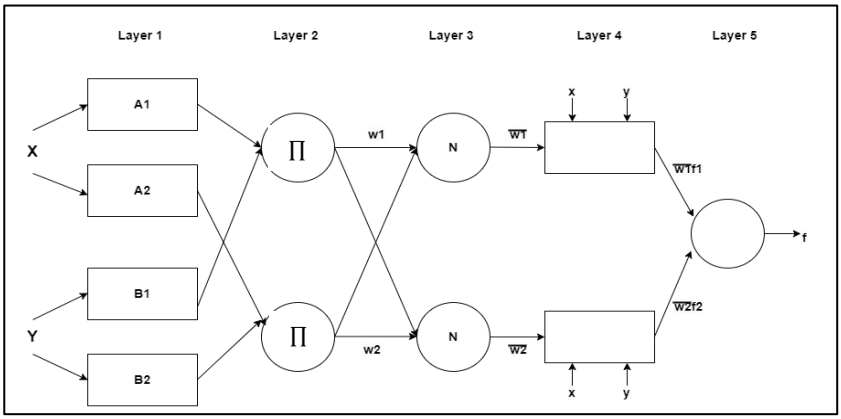

Figure 4 ANFIS model

1. If x is C_1 and y is D_1 then $f_1 = p_1x + q_1y + r_1$
2. If x is C_2 and y is D_2 then $f_2 = p_2x + q_2y + r_2$

Here C_i and D_i are fuzzy sets, x and y are inputs and the outputs within fuzzy regions are f_i defined by fuzzy rule. p_i , q_i and r_i are the design parameters whereas the i denotes the i^{th} rule (Übeylı & Güler, 2005). Figure 4 illustrates the ANFIS for 2 inputs resulting in one output with 2 rules. ANFIS is a 5-layer algorithm. O_i^j is the output of the i^{th} node and j^{th} layer.

- 1st layer: fuzzification of the input variables by membership function

$$O_i^1 = \mu_{A_1}(x), \qquad i = 1,2 \tag{15}$$

Here x would be the input and A_1 is the linguistic label. This label will be assigned by membership function.

- 2nd layer: application of the if then fuzzy rule

$$O_i^2 = w_i = \mu_{A_1}(x) * \mu_{B_1}(y), \quad i = 1,2 \tag{16}$$

- 3rd layer: computation of the output of each rule

$$O_i^3 = \overline{w_i} = \frac{w_i}{w_1 + w_2}, \quad i = 1,2 \tag{17}$$

- 4th layer: calculation of each rule based on their contribution

$$O_i^4 = \overline{w_i} * f_i = \overline{w_i}\,(p_i * x + q_i * y + r_i) \tag{18}$$

Here $\overline{w_i}$ is the output from 3rd layer.

- 5th layer: Crisp output obtained by defuzzification of the outputs

$$O_i^5 = \sum_i \overline{w_i} * f_i = \frac{\sum_i \overline{w_i} * f_i}{\sum_i \overline{w_i}} \tag{19}$$

Most of the time this classification has been used in the medical field of studies (Übeylı & Güler, 2005), and in the pattern recognition for the image classification of the medical conditions such as heart valve diseases, tumors etc. (Uğuz, 2012). Also used in chemical reactors (Lutfy et al., 2009), nuclear reactors (Man Gyun Na & Upadhyaya, 1998). As the fuzzy logic is being used widely in the data mining area; it was seen that the neural network will adapt to these scenarios.

An interesting study conducted by (Alparslan et al., 2011) focused on the data confidentiality. They created a hybrid method with the combination of ANFIS and SVM models aiming on the data protection of the military and government organization. The unclassified document was classified using ANFIS algorithm. Followed by stemming process where distinct Turkish terminologies were extracted (chi -square statics was made use for feature selection).

The ANFIS model used in this study uses the genfis3 which is an inbuilt MATLAB function used to generate the Sugeno type fuzzy inference system (FIS). It calculates the number of rules (fuzzy if then rules) necessary based on the input and output data being fed. On the other hand, the ANFIS function another in build MATLAB function implemented to train the ANFIS provides an

optimized membership function in order to minimize the difference between the actual and predicted output. All these processes are occurred in the 5-layer ANFIS network.

Weighted average will be the FIS overall output for all the rules applied in the model. The firing strengths of the rules is the weights. Hence the overall output would be the weighted average. The accuracy for each iteration with respect to number of features present also computed at the end for both types of fuzzy entropy measures.

4 Methodology

This section will be outlining the approach attempted to address the research questions : Finding the crucial factors affecting the vehicle CO_2 ratings, Classification models (fuzzy k-NN, Similarity classifier and ANFIS) that can be used by company in order to predict the CO_2 ratings in vehicles and how well the feature selection process influence in the performance of the classification models prediction of CO_2 of vehicles. The following areas are elucidated in this section process of data collection, pre-processing of data, feature selection using fuzzy entropy measures, development of model and evaluation. This structured approach gives a robust framework in order to identify and understand the significant forecaster of vehicle CO_2 permitting for more accurate and effective predictions.

4.1 Data Acquisition

The data set used for the study is from Kaggle (*2022 Fuel Consumption Guide,* 2022) which is publicly available, providing information about the fuel consumption of light duty vehicles in Canada in the year of 2022. With the intention of having a dataset which is current and pertinent.

This dataset contains 15 variables in total. Consisting of some key variables such as Engine size, number of cylinders, fuel consumption etc. whereas each of the variables will potentially impact the CO_2 emissions. The target variables in the dataset are CO_2 rating which are obtained from the vehicle's tailpipe. The ratings are scaled from 0 to 10 (worst to best). The data types of features which were considered are listed in Table 1.

Before proceeding with the analysis, the data was cleaned and preprocessed. In addition to this feature selection technique based on fuzzy entropy measures (De Luca & Termini, 1972) with similarity based classification procedure (Luukka & Leppälampi, 2006) , fuzzy kNN classification (Keller et al., 1985) and ANFIS classification (Jang, 1993) to find the best predictors for the task. It has been proved that statistical analysis is important when it comes to scientific studies (Ferreira, 2020). Henceforth in the following section the descriptive statistics of the data has been presented.

Categorical features	**Ordinal**	CO2 Rating
	Nominal	Vehicle Class Transmission Fuel Type Make Model
Continuous features		Fuel Consumption_City _L/100 km_ Fuel Consumption_Hwy _L/100 km Fuel Consumption_Comb _L/100 km Fuel Consumption_Comb _mpg Engine Size(L)
Discrete features		CO2 Emissions_g/km Cylinders

Table 1 Data types of the variables

4.2 Descriptive statistics

Summarization of data provided a high-level overview of the variables. The following table shows the descriptive statistics of each variable. It can be seen that the variable 'Model Year' has a variance of zero which will not provide useful information; hence it is removed from the dataset.

Variable	Minimum	Median	Maximum	Mean	Variance	Standard deviation
EngineSize_L	1.2	3	8	3.19	1.890	1.374
Cylinders	3	6	16	5.66	3.735	1.932
FuelConsumption_City_L_100Km	4	12.20	30.30	12.50	11.917	3.452
FuelConsumption_Hwy_L_100Km	3.90	9.200	20.90	9.36	5.221	2.285
FuelConsumption_Comb_L_100Km	4	10.80	26.10	11.09	8.273	2.876
FuelConsumption_Comb_mpg	11	26	71	27.24	59.06	7.685
CO2Emissions_g_km	94	257	608	259.1	4152.9	64.443
CO2Rating	1	5	10	4.53	2.166	1.471
SmogRating	1	5	7	4.95	2.821	1.6798

Table 2 Descriptive statistics

4.3 Data Preprocessing

Data preprocessing is considered to be first and foremost step in the data analysis process. The major issue of the inconsistencies and errors found in the data will be addressed in this process (García et al., 2015).

As an initial step the inspection for missing or inconsistent data was done and found that there are no such data available. The categorical data such as Vehicle Class, Transmission, Fuel Type, Model and Make were converted into numerical variables so that to make them usable in the models. Similarly, there were numerical features found in the dataset. The numerical features range from different scales. So, there is a tendency that large scale feature will overtake the low scale features creating a distorting difference in the performance of the model. hence it is a good practice to normalize the data ensuring the contribution of all the features equally (Aggarwal et al., 2001). Performed the min- max normalization (Shalabi et al., 2006) transforming the data in various scales into pre-defined interval of [0,1]. This method is considered to be used most frequently and the formula can be defined as,

$$x^{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (20)$$

Here x is the original dataset. x_{min} is the minimum value in the feature column and x_{max} the maximum value in the feature column, x^{scaled} is the normalized data point. Hence in the data set every minimum value of the feature will be converted to zero and maximum value transformed to 1. While the remaining values will attain a decimal value between 0 and 1. The class variables were not scaled or normalized as it is an unnecessary step when it comes to classification models (Graf et al., 2003).

Following the normalization, the outlier detection was performed. Outliers were detected using the interquartile range method. The 25% of the data will be the first quartile (Q1), that is the 25th percentile. Followed by the calculation of the third quartile (Q3), where the data lies at 75%. The interquartile range is computed as: IQR = Q3 – Q1. IQR is a statistical dispersion. Hence the outliers will be identified as those values lesser than Q1 – 1.5IQR and more than Q3 +1.5IQR. (in mathematical terms the lower bound and upper bound (Shalabi et al., 2006).

4.4 Data Sampling

The data is divided into two independent portions, as in training and testing data sets (Bramer, 2016). The training data sets are used to build the model while the testing data sets are those used for the evaluation of the performance of the model that is developed, testing data should not be included in the training or development of the model by any means (Witten et al., 2011).

The training and testing data are created randomly, in other words random sampling. In this point the data is prone to be separated unequally. This will lead to create unbalanced classes in the data (Kubat, 2017). As a solution for this kind of issue we can make use of the stratified sampling approach (Witten et al., 2011) which will make sure to divide the data in an appropriate manner concerning the class distribution that will in turn overcome any inaccurate results of the performance of the classification.

Once the data sampling is done in building the model, next process is the cross validation of the data. This is a widely used technique in order to evaluate the generalization while attenuating the overfitting problems (Ranganathan et al., 2019). There are different types of cross validation methods such as holdout, k-fold, leave one out cross validation etc.

Hold out cross validation is where some part of the data undergoes validation or else testing while the remaining data will undergo in training. This process is done repeatedly in order to avoid the sample bias and performance variance. With the final accuracy the performance measures and the classification accuracies will be averaged including all the iterations. The variance of the model can be gained. This method has been implemented in all 3 of the models used for this study.

An iteration of 30 runs was performed for all 3 models. The dataset is divided into training and testing with a ratio of 50:50. A stratified random sampling approach is used in every sampling step. The performance of the model in terms of accuracy and other measures are measured.

The test data which were excluded from the whole dataset is used to find the predictive performance of the fitted model. This shows that the validation data are used to tune the model parameters by minimizing the over fitting problems.

4.5 Performance Measures

Performance measures are used to evaluate the performance of the models. The predictions made on the training data are measured on the testing data. There are several methods to conduct this step. Confusion matrix for this particular step.

The number of correctly identified class samples : true positives (TP) , the number of correctly identified samples which are not included in the class: true negatives (TN), number of incorrectly identified samples : false positives (FP) and those not identified as class samples :false negative (FN) is how a classification model is evaluated based on the confusion matrix (Sokolova & Lapalme, 2009). In most cases this is used for the binary classification measures based on the confusion matrix values (Bramer, 2016). In the table 3 below the equations for the performance measures are listed (Nicolas & Peter, 2003).

Measure	Formula
Accuracy	$\frac{TP + TN}{TP + FP + TN + FN}$
Precision	$\frac{TP}{TP + FN}$
Recall / Sensitivity	$\frac{TP}{TP + FN}$
Specificity	$\frac{TN}{FP + TN}$

Table 3 Performance measure for binary classification.

When it comes to calculating the measures for multi class classification, we need to implement different way of calculation. That is a confusion matrix with C (>2) classes, this is defined by $\{a_{l,m}\}^S$ l = 1, m = 1, $a_{l,m}$ is an element of a l row and m column. The number of samples $l = m$, $a_{l,m}$ will be correctly classified. $l \neq m$ depicts the incorrectly classified samples of the class ω_l and ω_m. According to (Tharwat, 2021) the number of TP, TN, FP, FN are calculated as ,

$$TP(\omega_i) = a_{i,i} \tag{21}$$

$$TN(\omega_i) = \sum_{l=1}^{S} \sum_{l=1}^{S} (a_{l,m}) \quad (22)$$

$$FP(\omega_i) = \sum_{m=1, m \neq i}^{S} (a_{i,m}) \quad (23)$$

$$FN(\omega_i) = \sum_{l=1, l \neq i}^{S} (a_{l,i}) \quad (24)$$

By using these formulas, the sensitivity, specificity, precision for each class is calculated. It is stated that the final performance measure for a classifier is taken as the average of these performance measures. (Ferri et al., 2009).

Measure	Formula
Average Accuracy	$\frac{\sum_{i=1}^{l} \frac{TP_i + TN_i}{TP_i + FP_i + TN_i + FN_i}}{l}$
Specificity	$\frac{\sum_{i=1}^{l} TP_i}{\sum_{i=1}^{l} (TP_i + FP_i)}$
Sensitivity	$\frac{\sum_{i=1}^{l} \frac{TP_i}{TP_i + FN_i}}{l}$

Table 4 Performance classification for multi classification

Table 4 lists the generalized performance measure equations for multi classification. AUC is not calculated when it comes to the multi classification measures (Nicolas & Peter, 2003).

5 Results

5.1 Data

The dataset MY2022 Fuel Consumption Ratings is a multivariate dataset resourced from the Kaggle data science community. Following table 5 gives a clear understanding of the characteristics of the data.

No.of classes	No.of Features	No.of cases	Dataset characteristics
10	12	885	Multivariate

Table 5 Characteristics of data

5.2 Results and findings

As mentioned in the methodology the data was trained and tested after the feature selection was done. The following tables from 6 to 8 provides the results obtained from the 3 different models. The methods used is mentioned in the column1. That is the results with (De Luca & Termini, 1972) and fuzzy classifiers are given. In the 2nd row the results with (Parkash et al., 2008) and fuzzy classifiers are given. The column2 of the tables 6 to 8 states the mean accuracies obtained for the feature removal. The number of features used. The 4th column reports the features used. Followed by the variance obtained for the accuracies. Additionally, for the FkNN model the optimal number of k nearest neighbors is also mentioned. For the ANFIS model the optimal epochs have been recorded for the mentioned scenarios.

Method	Mean accuracies	No.of features	Features	Variance	Optimal k
FkNN	0.701	12	All	$4.2x10^{-5}$	6
FkNN + Luca	0.965	3	11,10,9	$2.3x10^{-5}$	1
FkNN + Park	0.981	2	11,10	$2.4x10^{-5}$	1

Table 6 Classification of FkNN model

Method	Mean accuracies	No.of features	Features	Variance
Sim	0.771	12	All	$2.2x10^{-4}$
Sim + Luca	0.953	1	11	$2.5x10^{-4}$
Sim + Park	0.935	3	11,10,9	$2.8x10^{-4}$

Table 7 Classification for Similarity based classification model

Method	Mean accuracies	No.of features	Features	Variance	Optimal Epoch
ANFIS	0.520	12	All	$2.5x10^{-2}$	100
ANFIS + Luca	0.970	2	11,10	$6.9x10^{-4}$	100
ANFIS + Park	0.972	3	11,10,9	$3.5x10^{-4}$	100

Table 8 Classification of ANFIS model

5.2.1 Results from FkNN model

The results for the FkNN model are reported in table 6**.** When only the similarity classifier was used a mean accuracy of 70% was achieved. Mean classification accuracy of 96% was achieved using the Luca feature extraction (De Luca & Termini, 1972) method with FkNN classifier. And a mean accuracy of 98% obtained with Park feature extraction (Parkash et al., 2008) method together with FkNN classifier. Out of all mean accuracies gained from these three scenarios it can be seen that by just using the FkNN classifier the accuracy was low. And when the Park extraction method was used it had the highest, with 2 features. It can be seen that with only 2 features the model was performing well, which in turn significantly reduced the processing time. The figure 5 shows how the accuracies varied along with the feature removal one by one in the dataset. Here only the Luca and Park feature extraction methods with FkNN classifier are plotted. It can be seen that the accuracies were considerably similar for both the methods having only slightest difference.

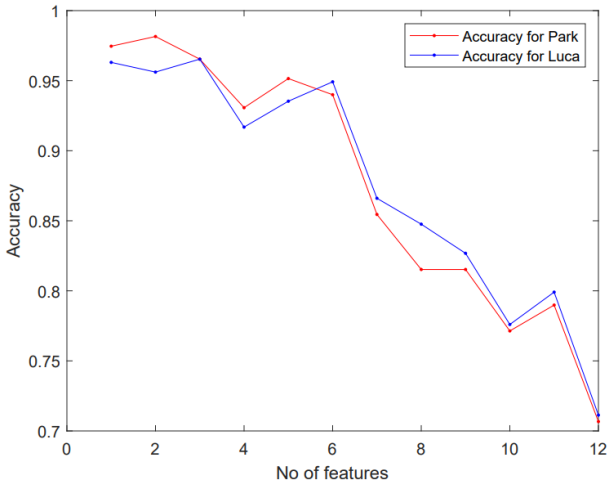

Figure 5 Mean accuracies of FkNN

Figure 6 and Figure 7 shows how the precision, sensitivity and specificity varies along with the feature removal in both Park and Luca methods. When it comes to the precision that is predicting the positive instances; both the methods show a downward trend. But Park + FkNN method outperforms the Luca + FkNN method. In a similar way the trend on sensitivity (recall) also tends to be the same where both methods are indicating a decline in the value as the number features

increased. Specificity showing a contrasting trend compared to the sensitivity and precision, remaining high for both the methods. Which means the negative instances are predicted accurately.

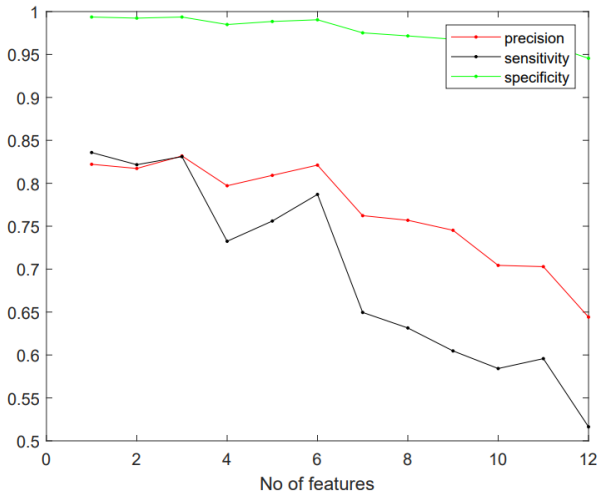

Figure 7 Performance measure FkNN + Luca

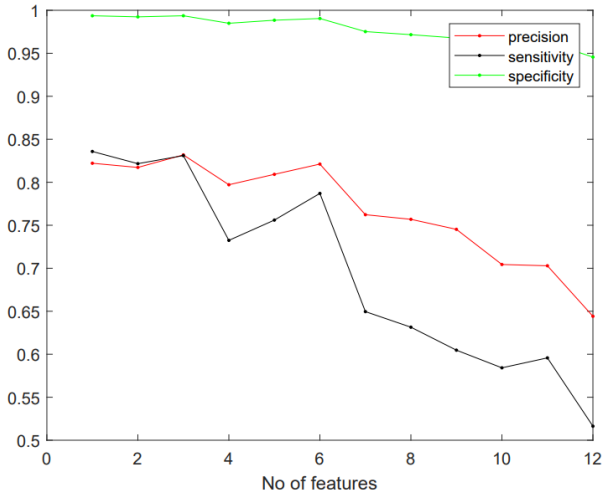

Figure 6 Performance measure FkNN + Park

5.2.2 Results from the Similarity classification

Table 7 shows the results from the similarity-based classification model. The mean accuracy is 77% with similarity classifier when all the features were present. Here when Luca feature extraction (De Luca & Termini, 1972) method together with similarity classifier was used the model's mean accuracy was 95% with only 1 feature. And the Park feature extraction method (Parkash et al., 2008) with similarity classifier achieved 93% of mean accuracy with 3 features. With 2% units the mean accuracy has dropped in this method compared to the Luca method. Figure 8 shows how the accuracies changed with the feature removal for both the feature extraction methods with similarity classifier. Even though the Luca + Sim method started with highest accuracy with 1 feature removal, compared to the Park + Sim method it was showing low accuracies when the features were removed. 65% of accuracy was achieved when 9 features were removed in Park +Sim method while Luca+ Sim 85% accuracy was obtained.

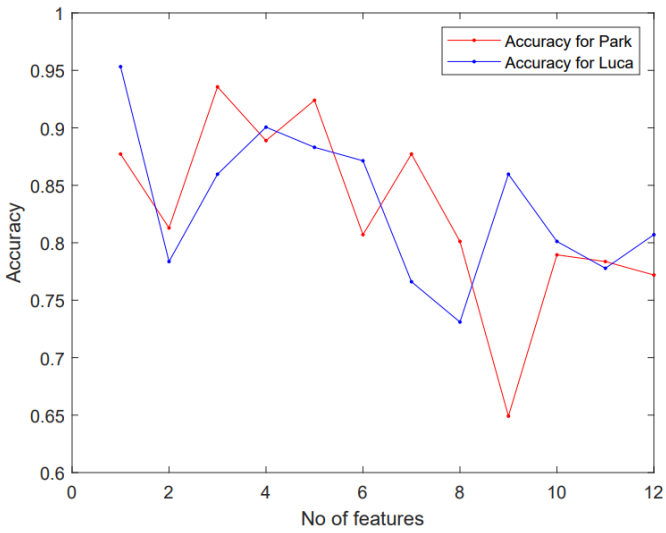

Figure 8 Mean accuracies of Similarity based classification

The performance measures are shown in the figure 9 and figure 10 The precision is significantly lower along the reduction of the features for both the Luca + Sim and Park + Sim methods. Sensitivity shows a lower trend in both the methods as well but compared to the Luca +Sim method shows remain stable while the Park +Sim method shows some variations. Across all the features both the methods are showing quite high values for the specificity ranging between 99 – 98%.

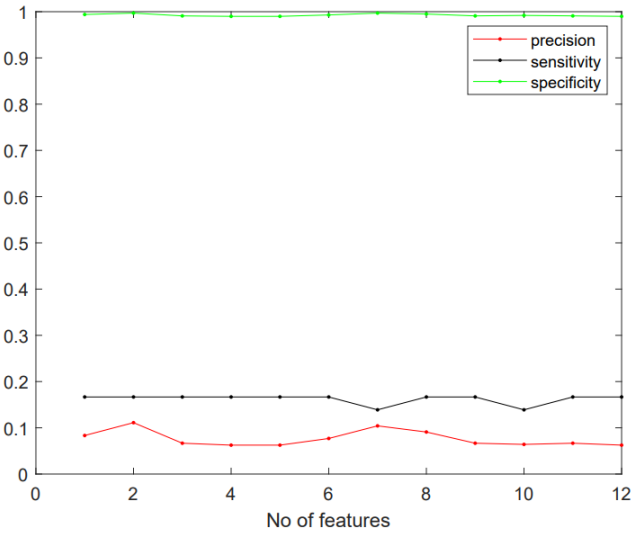

Figure 9 Performance measure Sim + Luca

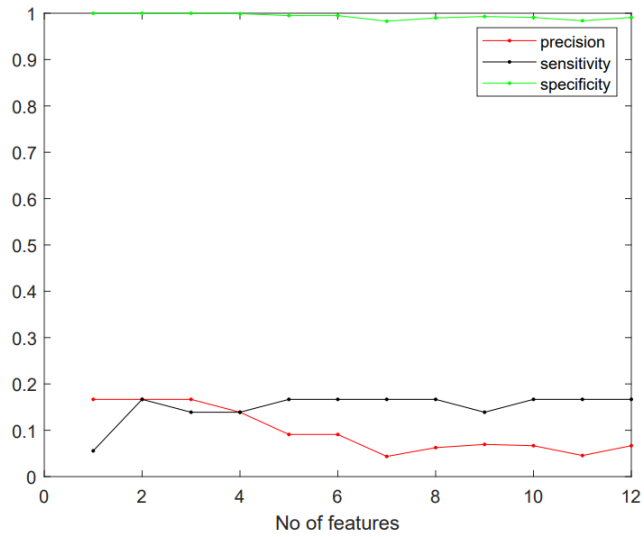

Figure 10 Performance measure Sim + Park

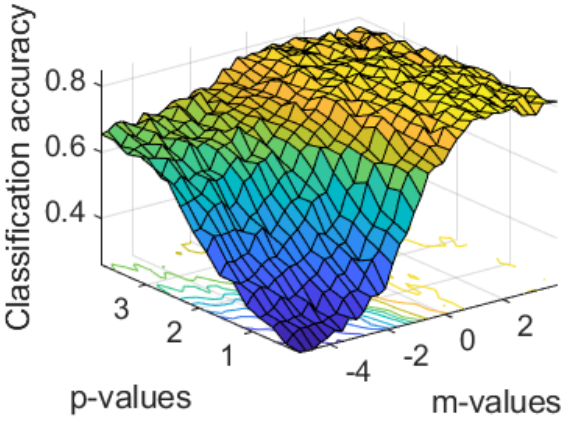

Figure 11 Parameters of Sim + Luca

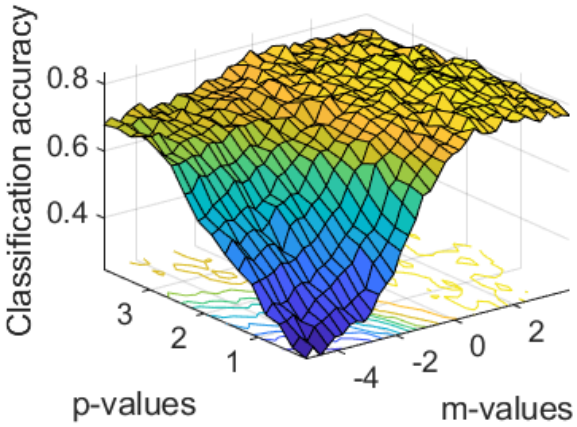

Figure 12 Parameters of Sim + Park

5.2.3 Results from the ANFIS classification

Results for the ANFIS classification model is listed in the table 8. The mean accuracy achieved by using similarity was 52%. Both the feature extraction methods Luca (De Luca & Termini, 1972) and Park (Parkash et al., 2008) together with ANFIS classifier gained approximately the same mean accuracy of 97% deviating only by 0.2%. This model achieved these optimal mean accuracies with only 2 features (Luca + ANFIS) and 3 features (Park + ANFIS). But it can be seen that the ANFIS classifier did not perform well, compared to the other two methods. Luca +ANFIS and Park +ANFIS behaved the same in the ANFIS model, in figure 13 it can be seen that the accuracies were more or less the same values along with the feature reduction. And gradually the accuracy was dropping when the number of features were high.

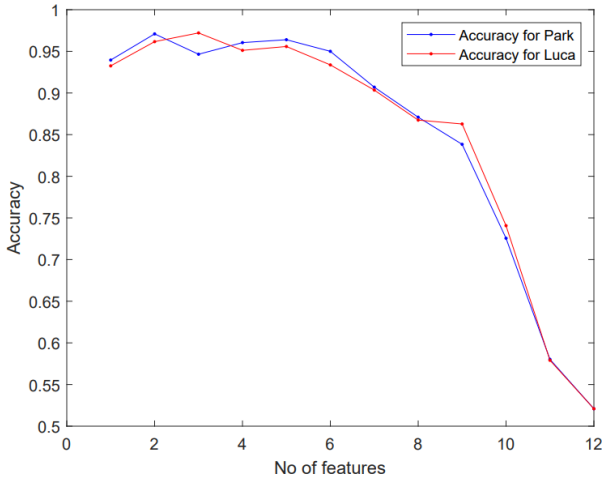

Figure 13 Mean accuracies of ANFIS

Figure 14 and Figure 15 and portray how the precision, specificity, and sensitivity of the ANFIS model varies with the feature removal. Precision values are high for both the scenarios, when the number of features is high the precision tend to reduce. But Park + ANFIS method represents a good trend compared to Luca +ANFIS method (having values between 98 – 93% from the beginning till the end). In a similar manner the sensitivity also indicates a high value. Here also Park + ANFIS method comparatively shows a high trend than Luca + ANFIS method. 52% of sensitivity was achieved with the presence of 12 features in Park + ANFIS method while it was 47% in the other method. Even the specificity range is high in both the methods. But Luca +ANFIS has shown a better result compared to Park +ANFIS in this performance measure. Where it shows a 99% with 3 features and 93% with 12 features, while for the other method it was 96% and 53% a drop in the results.

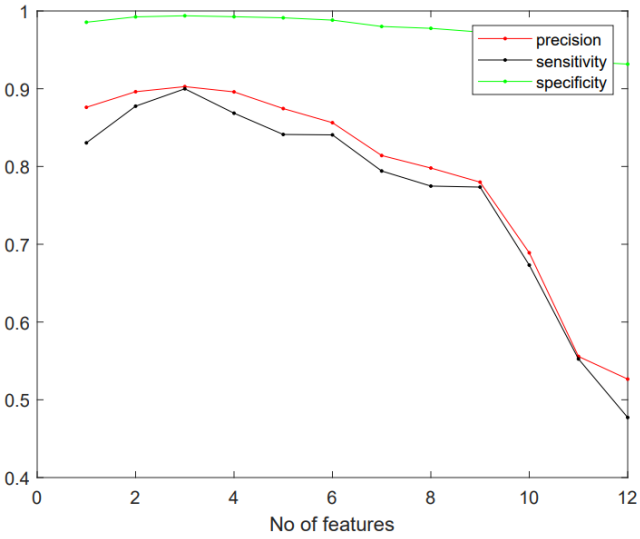

Figure 14 Performance measure ANFIS + Luca

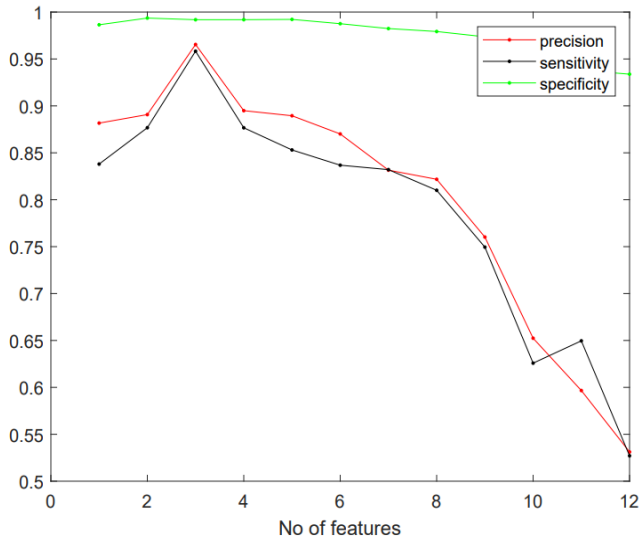

Figure 15 Performance measure ANFIS + Park

6 Discussion

This study explored the behavior of 3 different fuzzy classifiers using fuzzy entropy feature selection measures such as (De Luca & Termini, 1972) and (Parkash et al., 2008) in the context of CO2 ratings. In which the model performance was evaluated based on the presence of the fuzzy entropy-based feature selection. The impact of removing features were also analyzed based on the accuracy and stability of the classification models.

For the CO2 rating predictions, the FkNN classifier, Similarity based classification and ANFIS model performed well when the features were removed. When the features were reduced the models outperformed well achieving high accuracies with minimal variance showing the consistency of the results. This shows the feature removal allowed the models to identify the relevant features contributing to a better result. FkNN has been influenced the most by the fuzzy entropy measure compared to the other models as it yielded 98% of accuracy. There was a significant rise in the accuracy of ANFIS model when the feature selection methods were used, from 52% to 97%.

Moreover, the feature selection method indicates how well the fuzzy entropy measures have effectively improved the final outcome among the different machine learning models. This shows that this fuzzy entropy feature selection model can be applicable to various models and problems. As in a nutshell it can be said that from the outcomes of this research the importance of feature selection in machine learning is highlighted with the insights of the introducing various models.

7 Conclusion

The application of the feature selection method has significantly influenced the performance of the 3 classification models: FkNN, ANFIS and Similarity based classification model. It has shown a plausible result when coupled with FkNN and ANFIS models with substantiating levels of accuracies, precisions, sensitivities, and specificities, respectively for both CO2 ratings. It is often can be seen that the removal of the features leads to an improved model accuracy with reduced variance. But due to the model type and the type of the dataset this pattern may vary. It can be

From these predictions it can be concluded how well the feature selection techniques enhance the model performance. Also, the computational time is less. This can be pivotal to real world applications with bid data and complex feature dimensional spaces. It should be articulated that the discretion of feature selection method either with fuzzy entropy measures (De Luca or Park) depends on the type of classification model, the number of features and type of dataset. Future studies can be to explore on auxiliary feature selection methods and test their results on various machine learning models to further optimize performance and improve generalize the model's capabilities. As the dataset was struggling in class imbalance which was letting some of the classes not to be predictable. Working on to find out ways of overcoming imbalanced data for multi class datasets like this. For an example making use of the SMOTE or ADASYN methodologies to understand the classes.

References

2022 Fuel Consumption Guide. (2022). Government of Canada. https://natural-resources.canada.ca/sites/nrcan/files/oee/pdf/transportation/fuel-efficient-technologies/2022%20Fuel%20Consumption%20Guide.pdf

Aggarwal, C. C., Hinneburg, A., & Keim, D. A. (2001). On the Surprising Behavior of Distance Metrics in High Dimensional Space. In J. Van Den Bussche & V. Vianu (Eds.), *Database Theory—ICDT 2001* (Vol. 1973, pp. 420–434). Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-44503-X_27

Alparslan, E., Karahoca, A., & Bahşi, H. (2011). Classification of confidential documents by using adaptive neurofuzzy inference systems. *Procedia Computer Science*, *3*, 1412–1417. https://doi.org/10.1016/j.procs.2011.01.023

Baradwaj, B. K., & Pal, S. (2012). *Mining Educational Data to Analyze Students' Performance*. 7. https://doi.org/10.48550/ARXIV.1201.3417

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Blum, A. L., & Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artificial Intelligence*, *97*(1–2), 245–271. https://doi.org/10.1016/S0004-3702(97)00063-5

Bramer, M. (2016). Estimating the Predictive Accuracy of a Classifier. In M. Bramer, *Principles of Data Mining* (pp. 79–92). Springer London. https://doi.org/10.1007/978-1-4471-7307-6_7

Cai, D., Zhang, C., & He, X. (2010). Unsupervised feature selection for multi-cluster data. *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 333–342. https://doi.org/10.1145/1835804.1835848

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering*, *40*(1), 16–28. https://doi.org/10.1016/j.compeleceng.2013.11.024

Chen, H.-L., Yang, B., Wang, G., Liu, J., Xu, X., Wang, S.-J., & Liu, D.-Y. (2011). A novel bankruptcy prediction model based on an adaptive fuzzy k-nearest neighbor method. *Knowledge-Based Systems*, *24*(8), 1348–1359. https://doi.org/10.1016/j.knosys.2011.06.008

Dash, M., & Liu, H. (1997). Feature selection for classification. *Intelligent Data Analysis*, *1*(1–4), 131–156. https://doi.org/10.1016/S1088-467X(97)00008-5

De Luca, A., & Termini, S. (1972). A definition of a nonprobabilistic entropy in the setting of fuzzy sets theory. *Information and Control*, *20*(4), 301–312. https://doi.org/10.1016/S0019-9958(72)90199-4

Fernandez De Canete, J., Gonzalez-Perez, S., & Ramos-Diaz, J. C. (2012). Artificial neural networks for closed loop control of in silico and ad hoc type 1 diabetes. *Computer Methods and Programs in Biomedicine*, *106*(1), 55–66. https://doi.org/10.1016/j.cmpb.2011.11.006

Ferreira, Á. R. S. (2020). The importance of descriptive analysis. *Revista Do Colégio Brasileiro de Cirurgiões*, *47*, e20202682. https://doi.org/10.1590/0100-6991e-20202682

Ferri, C., Hernández-Orallo, J., & Modroiu, R. (2009). An experimental comparison of performance measures for classification. *Pattern Recognition Letters*, *30*(1), 27–38. https://doi.org/10.1016/j.patrec.2008.08.010

García, S., Luengo, J., & Herrera, F. (2015). *Data Preprocessing in Data Mining* (Vol. 72). Springer International Publishing. https://doi.org/10.1007/978-3-319-10247-4

Graf, A. B. A., Smola, A. J., & Borer, S. (2003). Classification in a normalized feature space using support vector machines. *IEEE Transactions on Neural Networks*, *14*(3), 597–605. https://doi.org/10.1109/TNN.2003.811708

Guyon, I., Elisseeff, A., & Kaelbling, L. P. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, *1*(3(7-8)), 1157–1182. https://doi.org/10.1162/153244303322753616

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning*, *46*(1/3), 389–422. https://doi.org/10.1023/A:1012487302797

Hassani, S. A., Oemisch, M., Balcarras, M., Westendorff, S., Ardid, S., Van Der Meer, M. A., Tiesinga, P., & Womelsdorf, T. (2017). A computational psychiatry approach identifies how alpha-2A noradrenergic agonist Guanfacine affects feature-based reinforcement learning in the macaque. *Scientific Reports*, *7*(1), 40606. https://doi.org/10.1038/srep40606

Huan Liu & Lei Yu. (2005). Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on Knowledge and Data Engineering*, *17*(4), 491–502. https://doi.org/10.1109/TKDE.2005.66

Hussein, A., & Thomas G, D. (1991). Learning with Many Irrelevant Features. *Department of Computer Science*, 547–552.

Jang, J.-S. R. (1993). ANFIS: Adaptive-network-based fuzzy inference system. *IEEE Transactions on Systems, Man, and Cybernetics*, *23*(3), 665–685. https://doi.org/10.1109/21.256541

Jinjie Huang, Yunze Cai, & Xiaoming Xu. (2006). A Wrapper for Feature Selection Based on Mutual Information. *18th International Conference on Pattern Recognition (ICPR'06)*, 618–621. https://doi.org/10.1109/ICPR.2006.198

John, G. H., Kohavi, R., & Pfleger, K. (1994). Irrelevant Features and the Subset Selection Problem. In *Machine Learning Proceedings 1994* (pp. 121–129). Elsevier. https://doi.org/10.1016/B978-1-55860-335-6.50023-4

Keller, J. M., Gray, M. R., & Givens, J. A. (1985). A fuzzy K-nearest neighbor algorithm. *IEEE Transactions on Systems, Man, and Cybernetics*, *SMC-15*(4), 580–585. https://doi.org/10.1109/TSMC.1985.6313426

Kira, K., & Rendell, L. A. (1992). A Practical Approach to Feature Selection. In *Machine Learning Proceedings 1992* (pp. 249–256). Elsevier. https://doi.org/10.1016/B978-1-55860-247-2.50037-1

Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, *97*(1–2), 273–324. https://doi.org/10.1016/S0004-3702(97)00043-X

Kubat, M. (2017). *An Introduction to Machine Learning*. Springer International Publishing. https://doi.org/10.1007/978-3-319-63913-0

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2018). Feature Selection: A Data Perspective. *ACM Computing Surveys*, *50*(6), 1–45. https://doi.org/10.1145/3136625

Lutfy, O. F., Noor, M. S. B., Marhaban, M. H., & Abbas, K. A. (2009). A genetically trained adaptive neuro-fuzzy inference system network utilized as a proportional-integral-derivative-like feedback controller for non-linear systems. *Proceedings of the Institution*

of Mechanical Engineers, Part I: Journal of Systems and Control Engineering, *223*(3), 309–321. https://doi.org/10.1243/09596518JSCE683

Luukka, P. (2011). Feature selection using fuzzy entropy measures with similarity classifier. *Expert Systems with Applications*, *38*(4), 4600–4607. https://doi.org/10.1016/j.eswa.2010.09.133

Luukka, P., Saastamoinen, K., & Kononen, V. (2001). A classifier based on the maximal fuzzy similarity in the generalized Lukasiewicz-structure. *10th IEEE International Conference on Fuzzy Systems. (Cat. No.01CH37297)*, *1*, 195–198. https://doi.org/10.1109/FUZZ.2001.1007281

Mailagaha Kumbure, M., Luukka, P., & Collan, M. (2020). A new fuzzy k-nearest neighbor classifier based on the Bonferroni mean. *Pattern Recognition Letters*, *140*, 172–178. https://doi.org/10.1016/j.patrec.2020.10.005

Maldonado, S., & Weber, R. (2009). A wrapper method for feature selection using Support Vector Machines. *Information Sciences*, *179*(13), 2208–2217. https://doi.org/10.1016/j.ins.2009.02.014

Man Gyun Na, & Upadhyaya, B. R. (1998). A neuro-fuzzy controller for axial power distribution an nuclear reactors. *IEEE Transactions on Nuclear Science*, *45*(1), 59–67. https://doi.org/10.1109/23.659555

Mehnaz Khan, & S. M. K. Quadri. (2013). Effects of Using Filter Based Feature Selection on the Performance of Machine Learners Using Different Datasets. *BIJIT - BVICAM's International Journal of Information Technology*, *5*(2).

Mitchell, T., & McGraw, H. (1997). *Machine Learning*. McGraw-Hill.

Modrzejewski, M. (1993). Feature selection using rough sets theory. In P. B. Brazdil (Ed.), *Machine Learning: ECML-93* (Vol. 667, pp. 213–226). Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-56602-3_138

Moore, A. W., & Lee, M. S. (1994). Efficient Algorithms for Minimizing Cross Validation Error. In *Machine Learning Proceedings 1994* (pp. 190–198). Elsevier. https://doi.org/10.1016/B978-1-55860-335-6.50031-3

Muhammad, I., & Yan, Z. (2015). SUPERVISED MACHINE LEARNING APPROACHES: A SURVEY. *ICTACT Journal on Soft Computing*, *05*(03), 946–952. https://doi.org/10.21917/ijsc.2015.0133

Nicolas, L., & Peter, F. (2003). *Improving accuracy and cost of two-class and multi-class probabilistic classifiers using ROC curves*. University of Bristol.

Parkash, O., Sharma, P., & Mahajan, R. (2008). New measures of weighted fuzzy entropy and their applications for the study of maximum weighted fuzzy entropy principle. *Information Sciences*, *178*(11), 2389–2395. https://doi.org/10.1016/j.ins.2007.12.003

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*(1), 81–106. https://doi.org/10.1007/BF00116251

Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. Morgan Kaufmann Publishers.

Ranganathan, S., Gribskov, M. R., Nakai, K., & Schönbach, C. (Eds.). (2019). *Encyclopedia of bioinformatics and computational biology*. Elsevier.

Rendell, L., & Seshu, R. (1990). Learning hard concepts through constructive induction: Framework and rationale. *Computational Intelligence*, *6*(4), 247–270. https://doi.org/10.1111/j.1467-8640.1990.tb00298.x

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*(6), 386–408. https://doi.org/10.1037/h0042519

Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, *23*(19), 2507–2517. https://doi.org/10.1093/bioinformatics/btm344

Santini, S., & Jain, R. (1999). Similarity measures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *21*(9), 871–883. https://doi.org/10.1109/34.790428

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, *2*(3), 160. https://doi.org/10.1007/s42979-021-00592-x

Shalabi, L. A., Shaaban, Z., & Kasasbeh, B. (2006). Data Mining: A Preprocessing Engine. *Journal of Computer Science*, *2*(9), 735–739. https://doi.org/10.3844/jcssp.2006.735.739

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, *45*(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Somol, P., Novovičová, J., & Pudil, P. (2006). Flexible-Hybrid Sequential Floating Search in Statistical Feature Selection. In D.-Y. Yeung, J. T. Kwok, A. Fred, F. Roli, & D. De Ridder (Eds.), *Structural, Syntactic, and Statistical Pattern Recognition* (Vol. 4109, pp. 632–639). Springer Berlin Heidelberg. https://doi.org/10.1007/11815921_69

Talpur, N., Salleh, M. N. M., & Hussain, K. (2017). An investigation of membership functions on performance of ANFIS for solving classification problems. *IOP Conference Series: Materials Science and Engineering*, *226*, 012103. https://doi.org/10.1088/1757-899X/226/1/012103

Tharwat, A. (2021). Classification assessment methods. *Applied Computing and Informatics*, *17*(1), 168–192. https://doi.org/10.1016/j.aci.2018.08.003

Übeylı, E. D., & Güler, İ. (2005). Automatic detection of erythemato-squamous diseases using adaptive neuro-fuzzy inference systems. *Computers in Biology and Medicine*, *35*(5), 421–433. https://doi.org/10.1016/j.compbiomed.2004.03.003

Uğuz, H. (2012). Adaptive neuro-fuzzy inference system for diagnosis of the heart valve diseases using wavelet transform with entropy. *Neural Computing and Applications*, *21*(7), 1617–1628. https://doi.org/10.1007/s00521-011-0610-x

Vommi, A. M., & Battula, T. K. (2023). A hybrid filter-wrapper feature selection using Fuzzy KNN based on Bonferroni mean for medical datasets classification: A COVID-19 case study. *Expert Systems with Applications*, *218*, 119612. https://doi.org/10.1016/j.eswa.2023.119612

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data mining: Practical machine learning tools and techniques* (3rd ed). Morgan Kaufmann.

Wolf, L., & Shashua, A. (2003). Feature selection for unsupervised and supervised inference: The emergence of sparsity in a weighted-based approach. *Proceedings Ninth IEEE International Conference on Computer Vision*, 378–384 vol.1. https://doi.org/10.1109/ICCV.2003.1238369

Yihua, C., E. K., G., Maya R, .Gupta, Rahimi, A., & Cazzant, L. (2009). Similarity-based Classification: Concepts and Algorithms. *US: MIT Press*, 30.

Zani, S., Cerioli, A., Riani, M., & Vichi, M. (Eds.). (2006). *Data Analysis, Classification and the Forward Search: Proceedings of the Meeting of the Classification and Data Analysis*

Group (CLADAG) of the Italian Statistical Society, University of Parma, June 6–8, 2005. Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-35978-8

Zhao, Z., & Liu, H. (2007). Semi-supervised Feature Selection via Spectral Analysis. *Proceedings of the 2007 SIAM International Conference on Data Mining*, 641–646. https://doi.org/10.1137/1.9781611972771.75