A comparison of Machine Learning algorithms for Classification and Timeseries Forecasting

Rohan Narayan Koli
MSc. Data Analytics
19224842

***Abstract*—Implementation of machine learning algorithms in various disciplines has become a need of the hour. Faster and accurate predictions or forecasts can be achieved using supervised machine learning. This paper tests the use of such algorithms in medical, financial and meteorological domains.**

In the medical domain, five classification algorithms are compared to predict the presence of cardiovascular disease. With an accuracy of 75.5%, Support vector machines with linear kernel has been suggested to outperform other classification algorithms. The models tuned for best hyper-parameters using k-fold cross-validation.

In the financial domain, three classification algorithms namely logistic regression, random forests and knn are compared to predict defaulting credit card customers. Random forest with an accuracy of 85.13% are well suited for classifying defaulting credit card customers.

In the third part, ARIMA models have been explored for accurate weather forecasts which produce highly significant results. This research will empower different parties such as medical professionals in diagnosing heart related diseases, banking professionals in customer profiling for funding needs and meteorologists in forecasting temperatures as well other related parameters.

***Index Terms*—Machine Learning, Logistic regression, Knn, Random Forest, ARIMA**

I. Introduction

In today's world machine learning and analytic has gained immense importance in our day to day lives. People demand faster, better and smarter results to various problems in their lives. To solve these problems efficiently, machines are being trained to make predictions of decisions. Behind the curtain's, a lot of data is been processed and numbers are being crunched by various algorithms to facilitate a desired solution.

According to WHO, "Cardiovascular diseases (CVDs) are the number 1 cause of death globally, taking an estimated 17.9 million lives each year"[1]. Our daily activities and habits such as routine exercise, smoking, alcohol consumption, dietary habits, stress, obesity, poor hygiene are the major risk factors for developing CDVs[2]. Since, CDVs are irreversible diseases, early detection along with swift diagnosis is the key in preventing deaths due to CDVs. Many lives can be saved if the diesase is detected beforehand as subjects can make necessary lifestyle changes and minimize the risk of CDVs. A simple blood sample test for testing cholesterol and blood sugar level coupled with blood pressure test can used to detect the probability of a CDV. In this paper, we attempt to train various classification algorithms to help predict the presence of CDV.

In the financial world, currency notes are being replaced by plastic money such as credit cards, debit cards and pre-paid cards. Out of these, credit cards provide many benefits such as a layer of security against fraudulent transactions, reward points, cashback schemes, discounts, sky miles programs and so on. In return, the banks that issue credit cards earn profit by getting annual fees, interest on delayed payments and other miscellaneous charges. But the system is not always foolproof. These banks run a major risk if some customers who have used all their credits and default. Recovering these credits can be cumbersome and may take many weeks, months or there may even be a possibility of no recovery at all. Hence, customer profiling, monitoring credit limits are essential for the business.By training machines to understand a pattern for defaulting customer will help minimize the risk of loss due to defaults. Our second objective in this research paper is to find the best suited machine learning algorithm to predict if the customer would default on his bill payments.

Accurate weather forecasts are critically important in the fields of agriculture, monitoring natural calamities and industrial resource/production optimization. Light conditions, soil and air moisture, rainfall are important parameters that affect the growth of crops. On the other hand, certain weather conditions encourage pests growth and diseases in crops. Also, crops cannot withstand strong winds and perish easily in such conditions. Secondly, natural phenomenon such as flooding, storms, wildfires, rise in sea levels can also be assessed by the weather conditions. Damage to people and property can be minimized by early forecasts of such disastrous situations. Lastly, industries such as mining, construction, energy, fishing and insurance are directly or indirectly affected by weather changes. The production cycles and ultimately their bottom lines of such industries are highly correlated to weather. Using machine learning techniques for forecasting we will be able to predict the temperature changes and rainfall which forms the third part of our third objective.

In the coming sections, we review the research that has been already conducted on the datasets. Next, we follow the knowledge discovering in databases (KDD) and data mining methodology pathway to evaluate the datasets by applying different machine learning algorithms to predict or forecast

[1]https://www.who.int/health-topics/cardiovascular-diseases

[2]https://www.mayoclinic.org/diseases-conditions/heart-disease/symptoms-causes/syc-20353118

data. Lastly, we evaluate and compare the performance of each model used and select the best suited model for the datasets.

II. Related work

Previously, similar work by numerous researchers has been conducted on the related datasets to diagnose cardiovascular disease using machine learning. In [1], Sandip Mondal discusses and compares various techniques such as support vector machine(svm), knn, random forest, naive bayes and artificial neural networks(ANN) for effectively identifying CDV's. The author suggests a hybrid model, Particle Swarm Optimization (PSO) in combination with SVM as the most effective classification model. The performance measures lack by not fine tuning knn and random forest algorithms fine tuning hyperparameters, not using AUC curve as an evaluation criteria, and lastly, by not splitting the predictor variable outputs in a stratified manner.

In another research [2], the author compares three machine learning algorithms, namely, Random forest, logistic regression and ANN and suggests logistic regression to be the best algorithm to classify the presence of heart disease in a patient. The performance parameters selected to compare the models were accuracy,sensitivity,specificity and kappa. The research lacked critical machine learning algorithms such as knn and naive bayes along with additional performance comparison parameters such as AUC. The data was randomly split into train and test in ratio 70:30 without considering stratified splitting. The parameters for random forest algorithm are not fine tuned to enhance performance.

In [3], the researchers predict the presence of CDV's using three machine learning algorithms namely, SVM, logistic regression and Framingham's risk model. On evalutaing, the non-linear model SVM with an AUC of 71% performs the best compared to other models suggesting, non-linear models outperform the linear ones. The advantage of SVM algorithm was explained efficiently by including grid search method to find the best model. In the method both linear and radial kernels and all possible combinations of parameters were considered. Restricting to limited selection of models and no evidence of stratified sampling are the shortcomings of the research.

In [4], the researchers, A. Maram, N. Mahalakshmi and N. Niriksha, compared six machine learning algorithms namely, naïve bayes, logistic regression, decision tree, random forest, SVM, multi-layer perceptron (neural network) and found neural networks to gain the highest accuracy of 91.2% followed by SVM and random forests (with accuracy 86.1%). Although ample number of algorithms are used, there is no evidence of using different types of SVM algorithms and lastly, no cross-validation techniques are used.

In relation with credit card default prediction, in [5] the authors, the researchers analyzed machine learning algorithms such as logistic regression, decision trees and random forest and concluded that random forests performed the best with an accuracy of 80%. On the positive front, a comparison of all models with the AUC is seen, but the research lacks cross validation techniques and hence, the parameters for the models are not fine-tuned.

In [6], the author Liyu Gui, ANN and gradient boosting algorithms along with random forest and logistic regression. ANN was cited to perform best among other algorithms with an accuracy of 81.6%. The author used a variety of algorithms. but the parameters were not tuned to get different models.

Furthermore, in [7], the researchers use six classification algorithms such as knn, Logistic regression,Discriminant Analysis,Naıve Bayesian, Neural networks and classification trees were considered.The researcher stresses ANN as the best performing algorithm with accuracy of 97%. For fine tuning only the ANN were tuned, which suggests lapses in comparison. In [8], the authors compared heuristic and machine learning approaches to predict the credit defaults. Random forest with an accuracy of 94% were well suited for predicting defaults. The research lacked appropriate exploratory analysis and hence data cleaning was not undertaken.

For weather prediction, A.A. Shafin in [9] used algorithms such as linear regression, polynomial regression, isotonic regression and Support Vector regressor and found isotonic regression to perform the best with the training set whereas support vector regressor performs best to predict the future temperatures.

In another study, [10], utilized, seasonal ARIMA, long short term memory to predict the forecast of Punjab,India. LSTM surpassed the performance compared with seasonal arima in forecasting the weather which is deep learning approach.

In [11], the research incorporates cyclic kernels, circular variables in regression trees, convolution neural networks (CNN) and lastly explored CNN with encoder-decoder networks which was best suited for weather forecasting.

In this research paper, all the with stratified sampling, we are consider fine tuning the parameters of random forest, knn and svm to evaluate each model critically. For weather forecasting, ARIMA modelling is considered with different seasonal parameters.

III. Methodology

In this research paper, Knowledge Discovery in Databases(KDD) methodology has been followed which helps in knowledge extraction from huge databases. The process of KDD is outlined below-

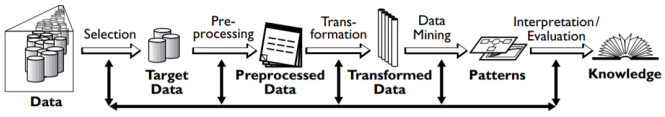

Fig. 1. KDD Methodology

The KDD process starts from data selection, and therefore the prior knowledge of the domain is necessary to generate a large database by selecting the predictors and outcome variable. The next step of data pre-processing involves data cleaning in which data is treated for missing values and noise if

any is removed. Data reduction step features identifying important factors and hence reducing the dimensionality which can be incorporated by using principal component analysis(PCA) or using functions (AIC) to include only important variables from the dataset.In the next step, data mining algorithms such as classification, regression or clustering are used to model using appropriate parameters. Patterns of data are identified which explains the data more correctly. The last step consists of interpreting the model performance, incorporating the knowledge acquired from models and reporting the models which can potentially be useful to parties [12].

A. Cardiovascular Disease

1) Data acquisition and pre-processing: The dataset contains medical records of 70,000 patients containing basic information along with habitual and lifestyle information over 12 variables or features. The target variable i.e the presence or absence of cardiovascular disease is binary coded. The dataset is acquired from kaggle and the link for the dataset is universally accesible from the below link-

https://www.kaggle.com/sulianova/cardiovascular-disease-dataset?select=cardio$_train.csv$

The dataset was checked for missing values using missmap() function from the library "Amelia" in R and no missing values were discovered. The age column initially scaled in days was converted to years for easy readability. Heights were selected in the range of 140 to 187 cm. Subjects having weights lower than 37 kgs were removed. Range of ap_hi was set between 90 to 250 and ap_low was set between 40 and 200. Levels of glucose and cholesterol were collapsed to binary.

Each relationship was checked with the outcome variable "presence of CDV" as in fig.2. On analysing, it was found that subjects aged between 50 to 60 years were more likely to have CDV. On categorizing on the basis of gender, it was found, Males to be more prone to CDV's. Subjects with weight 70-80 kgs, systolic blood pressure of 120mm, and diastolic pressure of 80mm, subjects with low outdoor activity, high presence of cholesterol, low glucose levels and active smokers had the most probability of presence of CDV's.

2) Research question: To use machine learning for accurate and automated diagnosis of presence Cardiovascular Diseases, given the age, gender, cholesterol levels, glucose levels, blood pressure, alcohol consumption and activity levels of a subject.

B. Credit Card Default

1) Data acquisition and pre-processing: The dataset is a credit card bill payment of a Taiwanese bank in the year 2005 of 30,000 customers with 23 variables. The variables include the customer's basic demographic profile description along with their past payment and billing records from April to September 2015. The data is available in an excel file format with no missing values.
archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients

The data was check for missing values and none were found. A correlation matrix between output variable and predictors suggested reasonable correlation between variables X1, X6, X7, X8, X9, X10 and X11 against Y. The data was split into train and test dataset with a ratio of 60:40 using stratified sampling as the output variable count that is default probability is low (22%). The levels of predictors X6 to X11 were collapsed as a part of pre-processing.

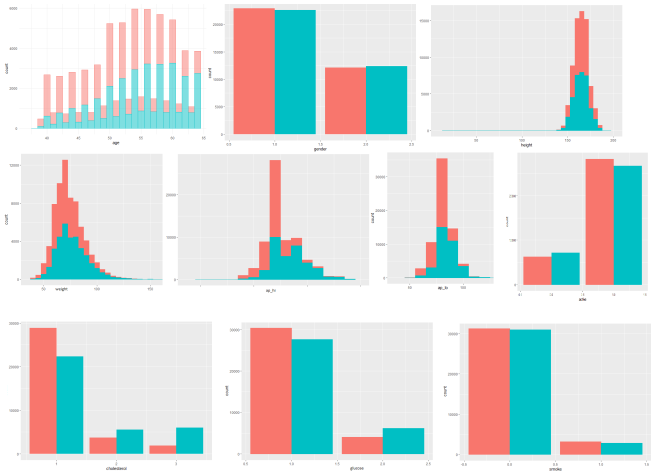

Fig. 2. CDV dataset

2) Research question: To predict defaulting customers before time using machine learning algorithms, and hence restrain further credits to the respective customers and ultimately minimize losses.

C. Climate Forecast

1) Data acquisition and pre-processing: The dataset contains a daily record of the climate parameters over 30 years collected from the *Dublin Airport* weather station of Ireland since January 01, 1980. There are 14,976 records along with 20 climate parameters variables. The data is available in a Comma-separated values file format with no missing values.
Source : https://www.met.ie/climate/available-data/historical-data

The data was split by 3:1 ratio. Data was converted to timeseries using ts() function. The time series was then decomposed to check seasonality, cycles and trends as showed in fig.3. We find the time series need On plotting the seasonal plot, we notice the temperature peaks in the months of June to September, whereas the months of December and January are the coldest.

2) Research question: To forecast daily temperature of Dublin,Ireland using historical daily temperatures using ARIMA forecasting technique.

D. Data mining algorithms selected

a) k-NN algorithm: K-NN algorithm is useful in the classification problems as it uses the nearest input in the train-set to determine the class of this nearest input [13]. The nearest neighbours are determined by the *Euclidean distance* or the shortest direct route method to determine the class [14]. Choosing the best value for 'K' is similar to maintaining the balance between bias and variance. Although this algorithm is

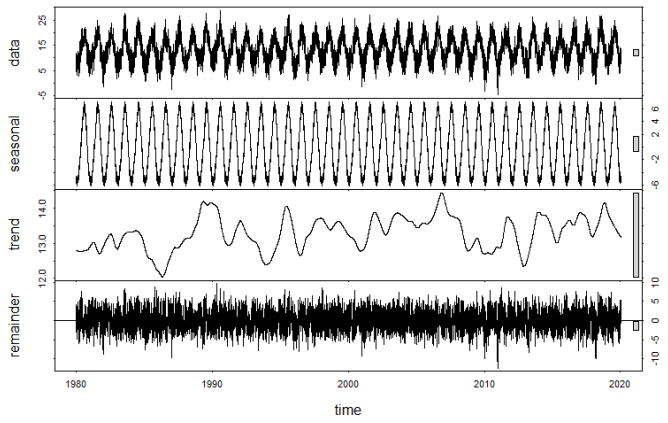

Fig. 3. Timeseries decomposition

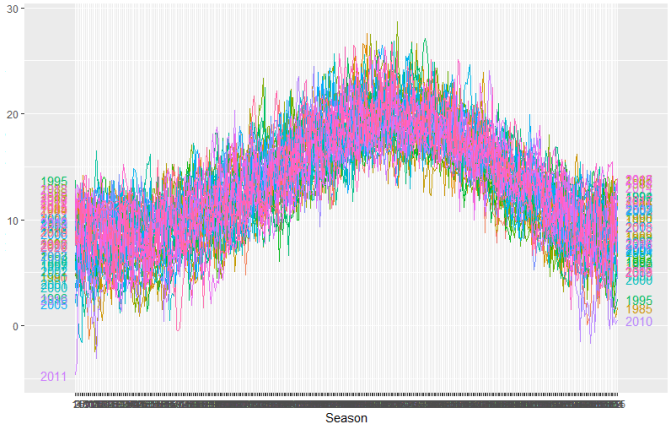

Fig. 4. Seasonal plot

categorized as *lazy* learning algorithm (since abstraction and generalization do not occur), it can be useful in predicting the probability of a person being diagnosed of Cardio vascular disease (CDV) from the dataset.

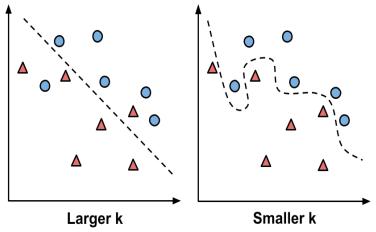

Fig. 5. knn

b) Naive Bayes: Naive Bayes is one of the popular and widely used classification algorithm that utilizes Bayes' conditional theorem as a base to classify the dependent variable. The algorithm assumes that all features in the dataset are equally important and independent; hence the word *naive* [15]. Although these assumptions are rarely true in the real world, this algorithm fetches fast and accurate results. This algorithm is therefore used in most real-time calculations such as in forecasting weather, medical science and stock markets. Being a classification problem to predict CDV, I am eager to test the effectiveness with the selected database.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

where, P(A—B) is the posterior probability of class given predictors; P(A) is the prior probability of class; P(B—A) is the the probability of predictor given class; P(B) is the prior probability of predictor.

c) Logistic regression: The dependent variable of our dataset (Default of payment) is a dichotomous variable and a problem of classification. Logistic regression incorporates a *logit* which is a natural log of the odds ratio and models the probabilities with two possible outcomes [13]. It has a 'S' or 'Sigmoid' shaped graph. Unlike linear regression, here the coefficients are estimated using maximum likelihood estimation (MLE).

$$P = \frac{1}{1 + e^{-(\beta_0 + \Sigma\beta_i X_i)}}$$

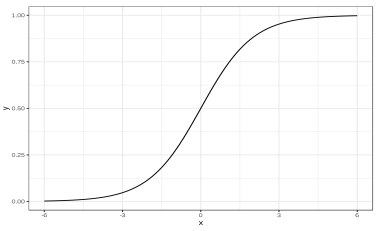

Fig. 6. Logistic Regression

d) Random Forest: Random forest is a step ahead of *Decision tree* model which when used selects bootstrapped data randomly from the dataset [14]. Next, the model selects parameters randomly as the root node forming a variety of individual decision trees. The output is generated by the *bagging* algorithm by aggregating the votes from each decision tree [15]. I chose Random forest as our problem question is of a *classification* type and the model handles both classification and regression tasks. The mechanism of the algorithm introduces randomness which generates more accurate results through cross-validation.

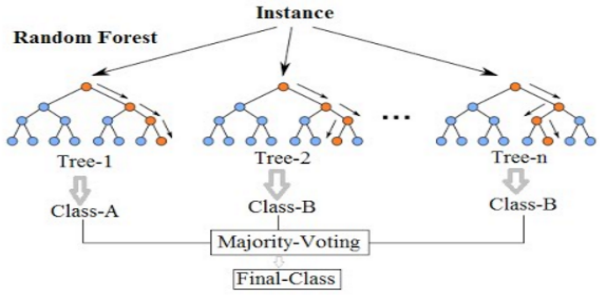

Fig. 7. Random Forest

e) ARIMA : Auto-Regressive Integrated Moving Average Model: Auto-regressive (AR) models predict values based on the past models whereas Moving Average models (MA) use only the error terms for forecasting. Combining both the techniques, we get the ARMA model. An ARMA model applied to d differenced data is called ARIMA. The model is based on the principle of *stationarity* which means that the mean and variance at any point in time should be stationary [16]. To check for stationarity, Augmented Dickey-Fuller (ADF) test is used. To make the data stationary, the model needs to be differenced. The lags are determined by Auto-corelation function (ACF) and Partial Auto-correlation function (PACF). Other than using multi-linear regression algorithm, I would also use the ARIMA algorithm to forecast temperature and compare the two models for accuracy.

f) Support Vector Machines (SVM): A support vector machine (SVM) falls under supervised machine learning model using classification algorithms for binary data. The main objective is to find a hyperp-lane in an N-dimensional space which critically classifies the data points.Maximization of the margin between the data points and the hyperplane is achieved using the loss function. For classification various kernels such as linear, radial and polynomial can be used.

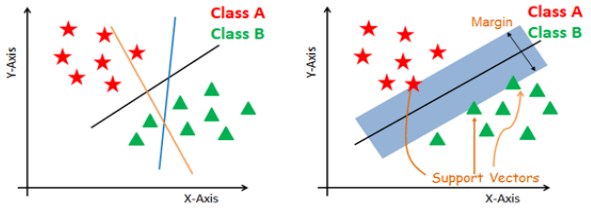

Fig. 8. SVM

IV. Model Evaluation

The following parameters have been considered for comparing different machine learning algorithms.

A. Model performance evaluation criteria

1) Confusion Matrix: The confusion matrix is a table which categorizes the predictions with their actual values.It is a square matrix whose dimension are equal to the predictors used [13]. The matrix classifies predictions into true positive (TP), true negative(TN), false positive (FP), false negative(FN). Using these categories, the model's accuracy and error rate can be measured as below:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Sensitivity = \frac{TP}{TP + FN}$$

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

Sensitivity and Specificity measures the positive and negative instances respectively that were correctly classified.

$$Precision = \frac{TP}{TP + FP} \quad \text{and} \quad Recall = \frac{TP}{TP + FN}$$

Precision measures how frequently, when a model makes a positive prediction, it ends up being right while Recall measures how certain we can be that all the occurrences with the positive predictions have been found by the model.

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

F1 measure is the harmonic mean of precision and recall which assumes values between 0 and 1; higher values indicate better performance.

2) ROC curve: The Receiver operating characteristic (ROC) depicts the performance of the classifier by plotting the True positive rate (TP Rate = 100 X TP/(TP+FN)) on vertical axis against the False positive rate (FP Rate = 100 X FP/(FP+TN)) on the horizontal axis [13].To create the curves, the predictions are sorted by the estimated probability of the positive class, in ascending order. Beginning at the origin, each prediction's impact on the TP rate and FP rate will result in a curve tracing vertically (for a correct prediction) or horizontally (for an incorrect prediction). The closer the curve is to the perfect classifier, the better is the model at predicting values which is measured with the help of area under the ROC curve (AUC).

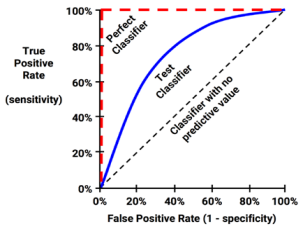

Fig. 9. ROC

B. Cardiovascular Disease

We compared six models using the CDV dataset namely, Logistic regression, Naive Bayes, Random Forest, k-nn and SVM. For applying the latter 4 models, the data was scaled and centered. We used 5 fold cross validation for these algorithms to achieve specific parameters. Using cross-validation, the value of k was set at 11 with highest accuracy of 71.06%. For Random Forest, the following parameters wer best suited-"mtry=2" , "best maximum nodes=33" and "best maximum trees =100".

Model Performance:

From fig.10.,the overall accuracy of knn classifier (75.55%) is seen best with the training dataset, whereas SVM with radial kernel performs the best with accuracy of 72.34%. Specificity is a critical parameter in the medical domain, as it measures the positive instances correctly classified, in our case, the presence of CDV correctly classified. Naive Bayes, has the

	Confusion Matrix							
	Train			Test			ROC	
Cardiovascular Dataset	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Train	Test
Logistic Regression	72.63%	78.16%	67.00%	71.90%	77.70%	65.98%	79.11%	78.22%
Naïve Bayes	71.97%	81.66%	62.11%	71.10%	81.37%	60.63%	78.79%	77.97%
Random Forest	72.77%	79.30%	66.10%	71.63%	78.87%	64.26%	78.12%	77.52%
Knn	75.55%	77.67%	73.40%	70.18%	72.71%	67.60%	83.49%	76.13%
SVM-Linear	72.52%	81.01%	63.88%	71.72%	80.73%	62.53%	-	-
SVM-Radial	73.31%	78.09%	68.44%	72.34%	77.55%	67.04%	-	-

Fig. 10. Model Performance - CDV

best specificity with both the training and test dataset, but the model is less accurate. Analysing the performance achieved by different models, we select SVM with Linear kernel model, as the model produces reasonable accuracy(average accuracy of 72%) along with second highest sensitivity(close to 81%).

C. Credit Default dataset

For predicting the credit card defaults, we used three machine learining algorithms: Logistic regression, Random forest and knn. For Random forest, mtry was set to 4, and ntree was set to 100. The value of k was set at 11 by 5 fold cross-validation. The summary of model comparison is shown in fig.11.

Model Performance: Random forest performs the best with

	Confusion Matrix							
	Train			Test			ROC	
Credit Default Dataset	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Train	Test
Logistic Regression	81.44%	94.65%	34.93%	81.41%	94.62%	34.90%	64.79%	64.76%
Random Forest	85.13%	94.71%	51.42%	80.50%	92.19%	39.35%	73.06%	65.77%
Knn	81.94%	94.44%	37.92%	81.05%	93.95%	35.65%	80.00%	73.77%

Fig. 11. Model Performance - Credit card default

train dataset whereas (accuracy 85.13%) whereas logistic regression performs the best with test dataset. Further comparing the AUC for both the models, we select Random Forest as the best algorithm with an AUC of 73% compared to logistic regression(AUC of 65%) for predicting defaults.

D. Weather dataset

For weather forecasting, ARIMA modelling is used. The time series was checked for stationarity using the adf.test() function and using ndiff(), nsdiff() function to checking if differencing is required to make the time series stationary. To make the timeseries stationary the series was seasonally differenced once and rechecked for stationarity which can be seen in fig.13. We get a significant p-value (less than 0.01) using the adf.test() suggesting no unitary root is present or the series is stationary.

		Train		Test	
Weather Dataset	AICc	RMSE	MAPE	RMSE	MAPE
Auto Arima (5,0,0)(0,1,0)[365]	52,138	2.78%	22.36%	3.61%	34.80%
ARIMA (2,0,1)(0,1,0)[365]	52,133	2.78%	22.37%	3.62%	34.76%

Fig. 12. Model Performance - Weather

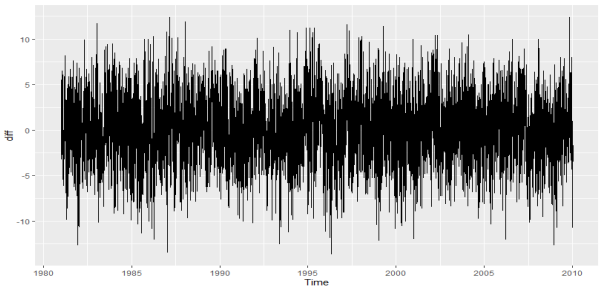

Fig. 13. Stationary series after differencing

We use the function auto.arima to get the best model. Arima (5,0,0)(0,1,0)[365] is suggested the best model by the function auto.arima. To investigate further we work with different models and find Arima(2,0,1)(0,1,0)[365] as a best performing model with minimum variables and low AICc with similar RMSE values. Lastly, performing Ljung–Box test on the residuls, we get a p value of 0.78, suggesting the residuals have no autocorrelations or residuals are white noise. Model Performance:

		Train		Test	
Weather Dataset	AICc	RMSE	MAPE	RMSE	MAPE
Auto Arima (5,0,0)(0,1,0)[365]	52,138	2.78%	22.36%	3.61%	34.80%
ARIMA (2,0,1)(0,1,0)[365]	52,133	2.78%	22.37%	3.62%	34.76%

Fig. 14. Model Performance - Weather

V. Conclusion and Future Work

This study has compared various classification models to classify presence of CDV's and defaulting credit card customers. For predicting CDV's, SVM worked the best. In the medical field, it is necessary to get a high sensitivity as the tests conducted could be lifesaving. Though SVM works the best in the dataset considered, the same needs to be cross-validated with other datasets in the medical field. Also, SVM's remain a black-box methods of prediction and identification and optimization of the parameters is left unsupervised. Diagnosing diseases using machine learning algorithms can reduce costs, help medical professionals with early diagnosis and ultimately save many lives. Therefore, the use of machine learning algorithms can lead to more accurate diagnosis of CDV's at an early stage, saving millions of lives.

In the second part, we compared three algorithms, logistic regression, random forests and knn to classify if a person will default. Random forest was well suited for the job with an accuracy of 85.13% with the training dataset. Utilizing such precise algorithms can help many bank and lending agencies to forecast the pattern of credit defaults and finally minimize huge unrecoverable losses. The model needs to be tested with other variables in the financial field to get a generalized model for predicting defaults. Hence, Random forest algorithms can be implemented when profiling customers in the banks and credits can be restricted beforehand.

Lastly, we compared different ARIMA models to forecast weather in dublin, Ireland. Such forecasts are necessary to mitigate to life and property due to natural calamities. The timing of harvest and using fertilizer is of utmost important in the field of agriculture which is completely dependent on the weather forecasts. Weather forecasts can also be useful in business operations by making available products that are dependent on weather. Importantly, weather forecasts help monitor global warming and keep a check on pollution levels.Hence, ARIMA model can be used for forecasting weather across the globe at different weather stations.

The modelling of algorithms is a data and time intensive process. The process demands high processing powers and huge data storage facilities. The research was limited to small datasets of nearly ten to fifty thousand observations. Also, the training time and cross validation for algorithms demanded processing times of couple of hours. This research was restricted to 5 fold cross validation due to huge amount of processing times required. Reproducibility and retraining the models require faster and more capable processing systems. Storing the results, also requires data warehousing. Hence, more time and resources would have resulted in more robust testing and tuning of hyper-parameters for the models yielding better results.

REFERENCES

[1] S. Mondal, "Diagnosis of cardiovascular diseases using hybrid feature selection and classification algorithms," Master's thesis, Dublin, National College of Ireland, December 2017. [Online]. Available: http://norma.ncirl.ie/3092/

[2] S. Ghosh, "Application of various data mining techniques to classify heart diseases," Master's thesis, Dublin, National College of Ireland, December 2017. [Online]. Available: http://norma.ncirl.ie/3082/

[3] P. Unnikrishnan, D. Kumar, S. Arjunan, H. Kumar, P. Mitchell, and R. Kawasaki, "Development of health parameter model for risk prediction of cvd using svm," *Computational and Mathematical Methods in Medicine*, vol. 2016, pp. 1–7, 08 2016.

[4] A. Maram, N. Mahalakshmi, and N. Niriksha, "Prediction of heart disease and diabetes using machine learning," 2020.

[5] Y. Sayjadah, I. A. T. Hashem, F. Alotaibi, and K. A. Kasmiran, "Credit card default prediction using machine learning techniques," in *2018 Fourth International Conference on Advances in Computing, Communication & Automation (ICACCA)*. IEEE, oct 2018. [Online]. Available: https://doi.org/10.11092Ficaccaf.2018.8776802

[6] L. Gui, "Application of machine learning algorithms in predicting credit card default payment," 2019. [Online]. Available: https://escholarship.org/uc/item/9zg7157qauthor

[7] I.-C. Yeh and C. hui Lien, "The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients," *Expert Systems with Applications*, vol. 36, no. 2, Part 1, pp. 2473 – 2480, 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417407006719

[8] S. R. Islam, W. Eberle, and S. K. Ghafoor, "Credit default mining using combined machine learning and heuristic approach," 2018.

[9] A. Shafin, "Machine learning approach to forecast average weather temperature of bangladesh global journal of computer science and technology: D neural artificial intelligence," 08 2019. [Online]. Available: https://computerresearch.org/index.php/computer/article/view/1858/1842

[10] P. Lohani, "Daily precipitation forecasting using neural network - a case study of punjab,india," 07 2019. [Online]. Available: http://norma.ncirl.ie/id/eprint/4324

[11] P. R. Larraondo, "Application of machine learning techniques to weather forecasting," 2019.

[12] G. P.-S. Usama Fayyad and P. Smyth, "The kdd process for extracting useful knowledge from volumes of data knowledge discovery in databases creates the context for developing the tools needed to control the flood of data facing organizations that depend on ever-growing databases of business, manufacturing, scientific, and personal information," vol. 39, 1996. [Online]. Available: https://sceweb.uhcl.edu/boetticher/ML$_{D}ataMining/p27-fayyad.pdf$

[13] B. Lantz, *Machine learning with R: learn how to use R to apply powerful machine learning methods and gain an insight into real-world applications*, 5th ed. Birmingham: Packt Publ., 2015.

[14] T. Hastie, R. Tibshirani, J. Friedman, R. Tibshirani, and J. Friedman, *Elements of Statistical Learning, The*.

[15] J. D. Kelleher, B. Mac Namee, and A. D'Arcy, *Fundamentals of machine learning for predictive data analytics*.

[16] R. H. Shumway and D. S. Stoffer, *Time series analysis and its applications*, 2017.