University of Cape Town

Course Code

STA 5076Z

Exploring Machine Learning Techniques: Support Vector Machines and Neural Networks

Author:
Khuliso Mmbi

Student Number:
MMBKHU001

May 23, 2024

Contents

Plagiarism Declaration

I, Khuliso Mmbi, hereby declare that the work on which this document is based on my original work (except where acknowledgements indicate otherwise) and that neither the whole work nor any part of it has been, is being, or is to be submitted for another degree in this or any other university. I authorize the University to reproduce for the purpose of research either the whole or any portion of the contents in any manner whatsoever.

Signature: Khuliso Mmbi
Date: 16 May 2024

1 Introduction

Machine learning is a subcategory of statistical learning which involves utilizing a wide variety of tools to understand data. This report will cover predictive modelling using Support Vector Machines (SVM) for classifications and Neural Networks(NN) for regression tasks. The two techniques will be evaluated on two data sets , the first one relating to heart failure clinical records and the second one will explore the number of bicycle share rentals per hour over the course of one year in Seoul, South Korea. The aim of this report is to build strong models that can reliably anticipate results based on pertinent features found in each of the datasets.

2 Support Vector Machines

SVMs are mainly used for classification, regression, and for detecting outliers. One of the advantages of the model is the ability to function effectively in high-dimensional spaces as a consequence of its capability to construct a decision boundary that maximizes the margin between the different classes. The upcoming sections will explore the SVM model on the heart failure dataset to predict the occurrence of death events given the explanatory variables.

2.1 Explanatory Data Analysis

	age	**creatinine_phosphokinase**	**ejection_fraction**	**platelets**	**serum_creatinine**	**serum_sodium**	**time**
Min.	40.00	23.0	14.00	25100	0.600	116.0	4.00
1st Qu.	52.00	109.8	30.00	210000	0.900	134.0	72.75
Median	60.00	247.5	38.00	262500	1.100	137.0	114.00
Mean	61.14	600.6	37.86	264226	1.415	136.7	130.87
3rd Qu.	70.00	582.0	45.00	305000	1.400	140.0	205.25
Max.	95.00	7861.0	70.00	850000	9.400	148.0	285.00

Table 1: Summary of Numerical Variables in the Heart Failure Dataset

Table 1 provides an extensive summary of the numerical variables in the heart failure dataset, including important statistical metrics such as the lowest, maximum, median, and quantiles values. For instance, creatinine phosphokinase values range from 23 to 7861 mcg/L and patient ages range from 40 to 95 years, with a median age of 60 years. Similarly, platelets counts vary from 25,100 to 850,000 kiloplatelets/mL, and ejection fraction percentages range from 14% to 70%, with a median of 38%. Serum sodium levels vary from 116 to 148 mEq/L, with a median of 137 mEq/L, while serum creatinine levels show a range of 0.6 to 9.4 mg/dL, with a typical value of 1.1 mg/dL. Furthermore, patients have a follow-up period which ranges from 4 to 285 days and a median of 114 days.

	anaemia	**high_blood_pressure**	**diabetes**	**sex**	**smoking**	**DEATH_EVENT**
0	133	159	140	80	161	163
1	107	81	100	160	79	77

Table 2: Summary of Categorical Variables in the Heart Failure Dataset

Table 2 summarizes all the categorical variables. The table shows that more patients do not certain conditions, such as anaemia(133), high blood pressure (159) and diabetes(140). In addition, the dataset is comprised of 80 females and 160 males, of which 161 patients are non-smokers and 79 are smokers. There are also 77 patients experiencing a death event and 163 surviving.

Modelling

2.2 Evaluation of Support Vector Machine using Radial Kernel Function

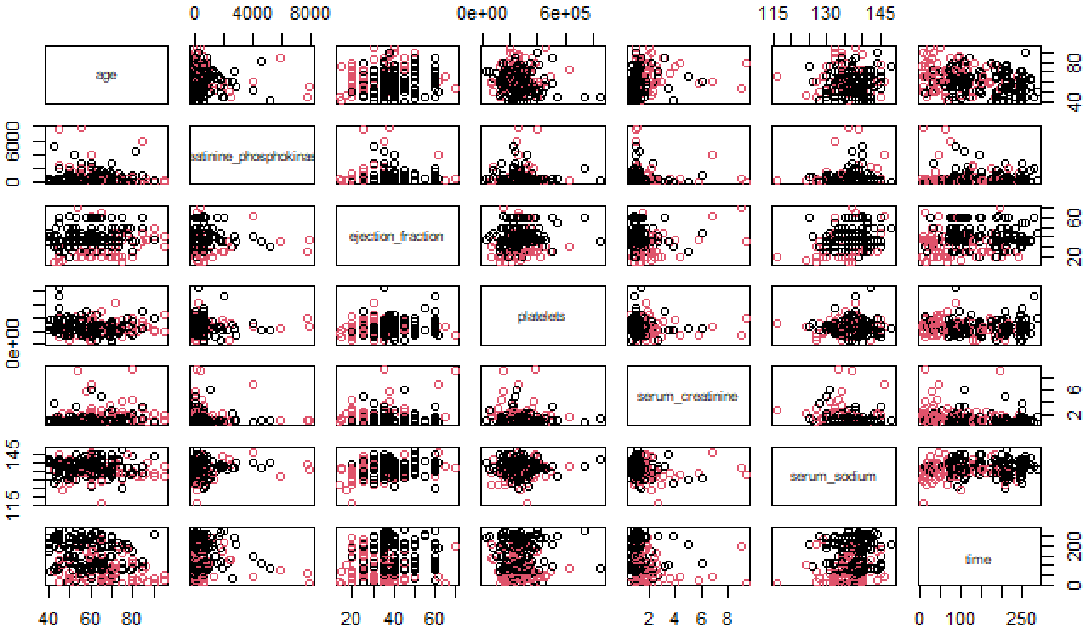

Figure 1: Relationship between different classes and features

Figure 2 displays the relationship between different classes and features as indicated by the red and black dots. There is visible overlap between the classes in most of the scatter plots which may indicate that the classes are not easily separable based on the features alone. Some scatter plots such as between age and serum creatine show a possible nonlinear trend.

To address this, the support vector machine with a radial kernel function is used. The kernel is ideal for this complexity as it effectively captures non-linear patterns in the data. The hyperparameters used are a cost parameter (C) of 0.1 and gamma of 0.1.

The results from the model can be seen in Table 3 below. From the training data, the accuracy of 83% ans on the test data, it slightly decreased to 81%. The recall was consistent on both data set with a value of 74%, this is a good indication that the model is bale to correctly predict positive outcomes on new data. Although the specificity decreased from 88% to 85%, this is still good as the model's ability to predict negative instances. The precision, F1-Score and Roc Auc were slightly higher in the training data than in the test, however the results still show that that the model can effectively differentiate between the classes when using the hyperparameters of a cost parameter (C) of 0.1 and gamma of 0.1.

	Train data	Test data
Accuracy	0.83	0.81
Recall	0.74	0.74
Specificity	0.88	0.85
Precision	0.74	0.70
F1 score	0.74	0.72
ROC AUC	0.93	0.87

Table 3: Comparison of evaluation metrics between Train data and Test data

2.3 Repeated Evaluation of Support Vector Machine with Radial Kernel Function

In order to assess the model's stability and generalisation performance, the SVM was simulated 100 times using different splits from the heart failure data set. The box plots below summarise the average of the performance metrics after the simulation. The box plot for accuracy and for specificity are close to 0.9 which suggests that the model performs well in terms of correctly classifying instances and avoiding false positives. Precision, Recall and F1-Score have a median of 0.8, which is not that different form the values obtained when the model was run once. The median for the ROC AUC is also almost above 0.8 .

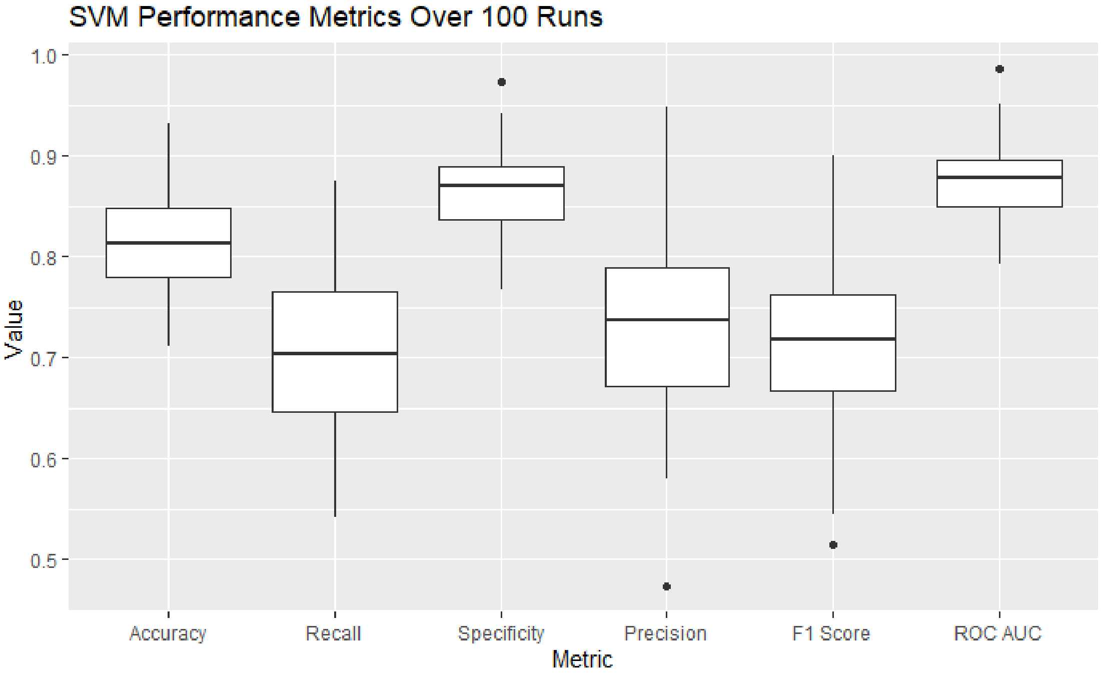

Figure 2: Performance metrics after simulation (averages)

2.4 Evaluation of Support Vector Machine with Different Cost and Gamma Parameters

Furthermore, to optimize the performance of the model, a systemic grid search was done using different values for each parameter. For cost (C), the values used are 0.1, 1, 10 and for gamma, the values used are 0.01, 0.1, 1. These values were chosen based on the results obtained from the initial model evaluation where a cost of 0.1 gave good results, so to test the models performance under different regularisation, high costs were used. The same approach was used for gamma where the model performed well initially with a lower gamma, but now to test for the models sensitivity, a higher gamma was implemented.

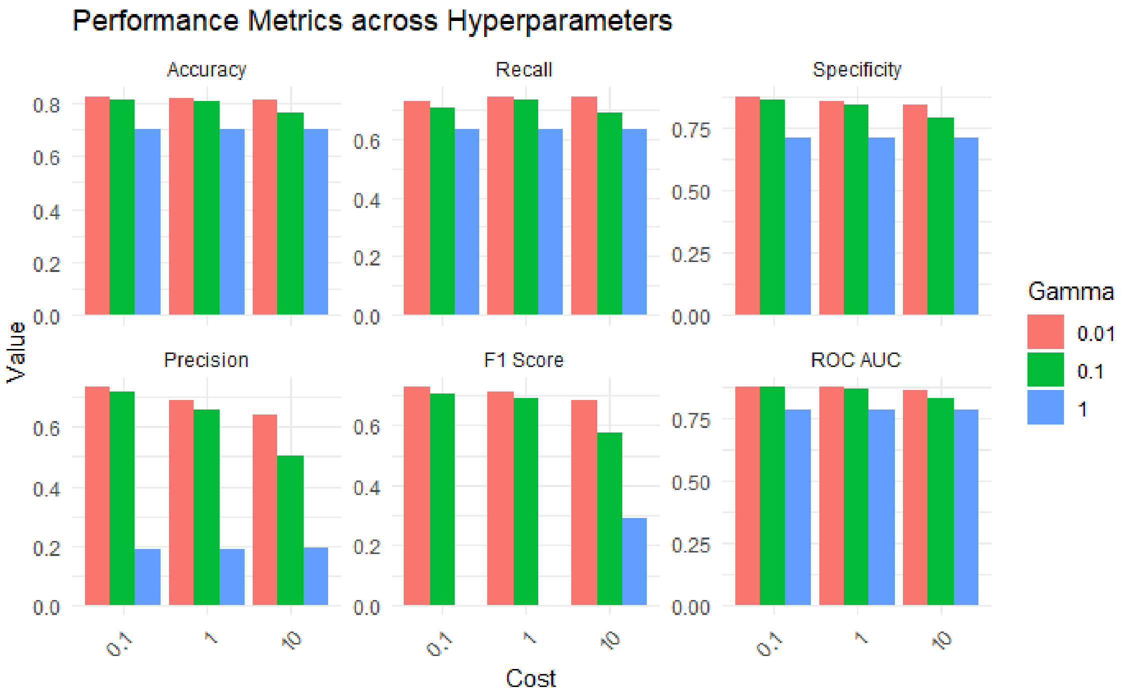

Figure 3: Performance metrics after simulation (averages)

Based on the results from the graphs in Figure 3, Accuracy, Recall, Specificity and the ROC AUC, all performed very well regardless of the cost and gamma used. Notably, a lower values for costs and lower values for gamma have the highest results. This shows that the overall model is capable of effectively differentiating between positove and negative instances. In contrast, Precision and F1 Score did not perform well will a high values of cost and gamma.This could suggest that the model cannot find a balance between precision and recall when cost and gamma are high. This can further lead to a high number of false postitives as can be seen by the low graphs for precison and F1 score.

2.5 Hyperparameter Tuning and Evaluation using Grid Search

As an attempt to optimize the performance of the model, the hyper-parameters were fine-tuned using grid-search based on the initial evaluation. A performance grid search was employed to to ensure reliable evaluation across iterations. The best parameters after running the model were at a cost of 0.0509 and gamma of 0.025. The model was reevaluated, showcasting improvement as can be seen by the table 3 below. The model's accuracy has now improved to 89.83%, recall improved to 87.50%, precision of 97.22%, Specificity of 97.22%, F1 score of 92.11%, and ROC AUC of 97.89 These results are all higher than for the model in section 2.2 which shows the importance of parameter tuning.

Metric	**Value**
Accuracy	0.898
Recall	0.875
Precision	0.972
Specificity	0.972
F1 Score	0.921
ROC AUC	0.979

Table 4: Performance Metrics on Test Data

2.6 Comparison of Grid Search and Repeated Runs

Metric	**Repeated runs**	**Grid Search**
Accuracy	0.809	0.898
Recall	0.706	0.875
Specificity	0.866	0.972
Precision	0.719	0.972
F1 Score	0.708	0.921
ROC AUC	0.875	0.979

Table 5: Comparison of performance metrics between repeated runs and grid search.

Table 5 shows the performance metrics for the model constructed through grid search and the simulated model. It is evident that the grid search outperforms repeated runs in terms of the model's performance metrics, including recall, specificity, accuracy, F1 score, and ROC AUC. For instance, the recall increased from 70.6% in repeated runs to 87.5% when grid search was used, and the ROC AUC increased from 87.5% to 97.9%. Compared to repeated runs, these variances imply that grid search was a more efficient method of optimizing the model's performance.

2.7 Summary

In conclusion, SVMs proved to be efficient in predicting hear failure events. It was able to capture some of the complex and non linear relationship. The results highlight the impact of hyperparmeter fine-tuning through grid search, on the performance of the svm model. Initially, when the process was first done with a cost of 0.1 and gamma of 0.1, the results were decent, however, the results started to improve with parameter adjustments, in our case it was as a result of low cost that was found at 0.05 and a sightly higher gamma of 0.25. These results are evidence that when optimized correctly, SVM can be a useful tool for classifying and predicting heart failure events.

3 Neural Networks

Neural Networks are another type of machine learning techniques that was explored on this paper. NNs are modelled after the operations of a human brain, comprised of layered networks of connected nodes, also referred to as neurons. The NNs are more advantageous due to the ability to modify the weights between neurons to discover complex patterns and relationship in the data.

The upcoming section will use neural networks to predict the number of bicycle rentals per hour over the course of one year in Seoul, South Korea explored using different parameters.

3.1 Explanatory Data Analysis

In order to test the model's prediction accuracy, the data set was split into a training set and testing set using an 80-20 split.

Table 6 summarizes the correlation matrix between bike rentals and all predictor variables. From the table, the strong predictors with positive correlations are Temperature, Hour, Dew Point Temperature, and Dew. The weak predictors are Wind Speed, Visibility, Solar Radiation, Rainfall, and Snowfall. There is some multicollinearity between some variables, such as Temperature and Dew Point Temperature, as indicated by their high correlation of 0.91. The metrics will be important when selecting features for the neural network model.

	Rented Count	Hour	Temp	Humidity	Wind Speed	Visibility	Dew Point	Solar Rad	Rainfall	Snowfall
Rented Count	1.00	0.41	0.54	-0.20	0.12	0.20	0.38	0.26	-0.12	-0.14
Hour	0.41	1.00	0.12	-0.24	0.29	0.10	0.00	0.15	0.01	-0.02
Temp	0.54	0.12	1.00	0.16	-0.04	0.03	0.91	0.35	0.05	-0.22
Humidity	-0.20	-0.24	0.16	1.00	-0.34	-0.54	0.54	-0.46	0.24	0.11
Wind Speed	0.12	0.29	-0.04	-0.34	1.00	0.17	-0.18	0.33	-0.02	-0.00
Visibility	0.20	0.10	0.03	-0.54	0.17	1.00	-0.18	0.15	-0.17	-0.12
Dew Point	0.38	0.00	0.91	0.54	-0.18	-0.18	1.00	0.09	0.13	-0.15
Solar Rad	0.26	0.15	0.35	-0.46	0.33	0.15	0.09	1.00	-0.07	-0.07
Rainfall	-0.12	0.01	0.05	0.24	-0.02	-0.17	0.13	-0.07	1.00	0.01
Snowfall	-0.14	-0.02	-0.22	0.11	-0.00	-0.12	-0.15	-0.07	0.01	1.00

Table 6: Correlation Matrix

Table 7 explores the categorical variables, like what season it was or if it was a holiday. This table reveals that most bike rentals happened in the spring and on holidays. It also shows that most of the days when bikes were rented were regular functioning days and not special occasions. This information gives us a basic idea of how the data is structured, which will be helpful for further analysis.

Variable	**Frequency**
Seasons	
Autumn	1741
Spring	1778
Summer	1764
Winter	1726
Holiday	
Holiday	343
No Holiday	6666
FunctioningDay	
No	231
Yes	6778

Table 7: Summary Statistics of Categorical Variables

3.2 Neural Network Construction and Performance Evaluation

The neural network (NN) for this section has been built with 1 hidden layer. To avoid overfitting, L1 (Lasso) regularization with a coefficient of 0.0001 was applied to the weights of the hidden layer.

Parameters employed:

Loss function: Default Mean Squared Error (MSE)

Optimiser: Adaptive learning rate method (H2O package)

Epochs: 1000

Bach size: Automatically determined by the H20 function in RStudio

Table 8 below shows the Root Mean Squared Error (RMSE) metrics for the training data and the test data. The RMSE measures the average difference between the predicted values and the actual values of the model. From the table, the RMSE for the training data is 386.681 and the test data is 367.530 which is slightly lower, this is a good indication that the model is not over-fitting on the test data is able to perform just as well on data that is new.

	Training Data	**Test Data**
RMSE	386.681	367.530

Table 8: RMSE values for Train Data and Test Data

Further more, figure 4 shows the neural network graph which was constructed using 1 hidden layer with 2 neutrons. The input layer is made up of 17 nodes corresponding to the features: Date, Hour, Temperature, Humidity, Wind speed, Visibility, Dew point temperature, Solar Radiation, Rainfall, Snowfall, and the categorical variables for Seasons(Autumn, Spring, Summer, Winter), Holiday (No Holiday), and Functioning Day (Yes). The hidden layer has 2 neurons, with weights denoted by the lines. For example, the weight from the Humidity feature to the first neuron is 0.08933, indicating a moderate positive influence, while the weight from Solar Radiation to the same neuron is 0.07002. The output layer has a single node representing the predicted count of rented bikes, with weights of 2.25682 from the first neuron and -1.18584 from the second neuron, illustrating their respective positive and negative contributions to the prediction.

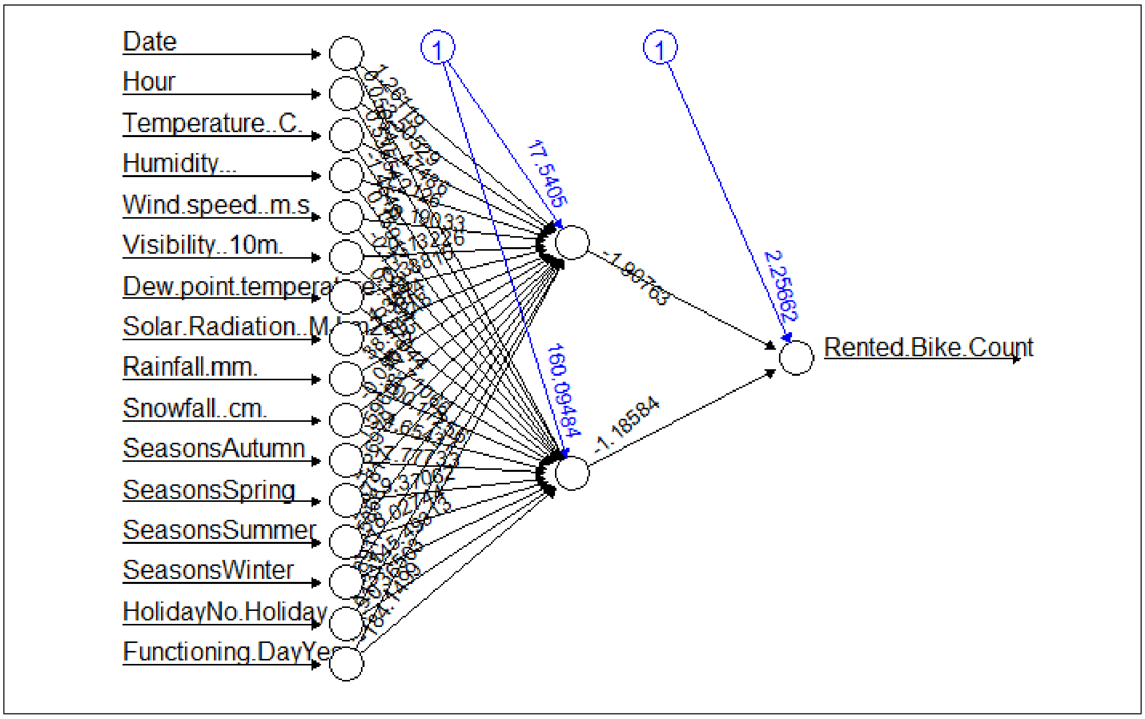

Figure 4: Neural Network plot

3.3 Model Optimization and Parameter Tuning

The model was further explored using different hidden layers, neurons and regularization rate. This was achieve through the use of grid search to find the best parameters. Grid search tests the different hyper-parameter combinations that best optimize the model. Cross validation was also employed to determine the best parameters.

Table 9 summarises the results from parameter tuning. The best model had an RMSE of 343.971 on the training data. This is a good indication of performance in predicting the number of bicycle rentals. In addition, this is the highest as compared to other models explored in 3.2 reflecting the power of grid search.

Parameter	Value
Hidden Layers	3
Neurons	15
L1 Regularization Rate	0.001

Table 9: Best Model Parameters

3.4 Best Model Performance and variable importance

As a result of an increase in more hidden layers, neurons and regularization, the best model's complexity is larger than the firts model in 3.2. The advantage of this approach, is it can be able to capture even some of the complex relationship that were not captured in the first model, however there is also a risk of overfitting.

The best model achieve an RMSE of 330.721 on the test data which is lower than that from the training data. This improvement demonstrates the mdels predictive ability compared to the simpler model.

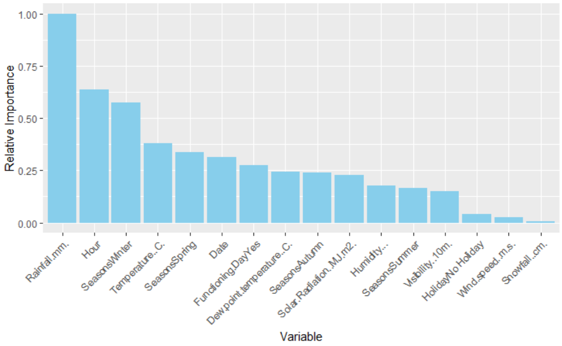

Figure 5: Variable importance for Model 1

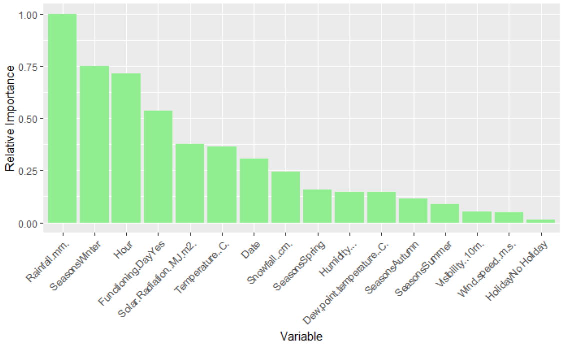

Figure 6: Variable importance for Best Model

Variable importance comparison

Figure 5: Model 1 (Baseline model) Top variables: Rainfall, Hour, Season(Winter), Temperature The top variables Rainfall and Hour suggest that the weather conditions and the time are the most important determinants of bicycle rentals.

Figure 6: Best Model Top variables: Rainfall, Season(Winter), Hour, FunctionalDay Similar to model 1, the weather conditions plays a big factor in bicycle rentals, however, for the best mode, there is a shift in the importance ranking, with the Season and Functional day being one of the top variables.

Both models shows Rainfall and Hour as the most important variable for predicting bicycle rentals. The improved performance of the best model as seen by the decrease in RMSE displays its improved capability to effectively predicting the bicycle rentals when using these key variables.

3.5 Summary

The neural networks proved to be efficient in their ability to uncover complex data. The most important variables were identified to be Rainfall, Hour, Season (Winter), and FunctionalDay. There was also an improvement in RMSE from 343.971 to 330.7211 from Model 1 to the best model, highlighting the importance of optimizing parameters.

These improvements further indicate the significance of fine-tuning neural networks for predicting the number of bicycles rented based on different variables. With proper optimization, neural networks can play a crucial role in making accurate bicycle rental predictions, thereby supporting better decision-making.

4 Conclusion

In conclusion, both the SVM and NN models proved to be valuable machine learning techniques when appropriately tuned. The SVM model effectively improved the classification of heart failure events, demonstrating its capability in handling high-dimensional classification tasks. Meanwhile, the NN model successfully uncovered complex relationships and identified key features that contribute to the number of bicycles rented. Both models highlighted the critical importance of parameter optimization in enhancing model performance, revealing the importance of fine-tuning to achieve reliable and accurate predictions.

5 References

1. James, G., Witten, D., Hastie, T., Tibshirani, R. (Year). *An Introduction to Statistical Learning with Applications in R* (2nd ed.). Springer.