Introduction

Our group used the Auto MPG (Quinlan, 1993) dataset for the regression task and Maternal Health Risk (MHR) (Ahmed, 2020) for the classification task. For each task we implemented linear and logistic regression, support vector machine (SVM), decision tree (including random forest), and multi-layer perceptron (MLP) neural network machine learning methods. We applied K-fold cross-validation on the whole dataset to optimise our models by tuning the hyperparameters. We evaluated model performance on 30% of the dataset, using classification accuracy (CA) and F1 score for the classification task and mean squared error (MSE) and root mean squared error (RMSE) for the regression task.

Data Analysis & Cleaning

Classification Dataset (MHR)

The MHR dataset contained 1013 instances and six features: age; systolic blood pressure (sBP); diastolic blood pressure (dBP); blood glucose; body temperature and heart rate (HR), to predict maternal risk as low, medium, or high.

We decided not to apply Principal Component Analysis (PCA) to reduce dimensionality despite high correlation between sBP and dBP (R=0.79) because the data already has few features, and we wanted to retain as much information for our models which is particularly important in a medical setting.

The dataset was well-balanced across the classes, with a small number of extreme outliers in HR (below 10 bpm) so we decided to remove these as faulty samples. While half the dataset are potential duplicates, we kept the samples as the class distribution was unaffected by their removal.

Regression Dataset (Auto MPG)

The Auto MPG dataset contained 398 instances and seven features: displacement, cylinders, horsepower, weight, acceleration, model year and origin. We removed origin because the feature meaning was ambiguous. We imputed missing values in horsepower with the median. We kept the outliers in horsepower and acceleration because they did not deviate significantly from the overall distribution.

Due to different variances within the feature values, we implemented zero-mean normalisation to prevent any single feature dominating the models. Displacement, horsepower, and weight were highly correlated (Figure 1); however, we chose not to reduce their dimensions through PCA because these were the top-ranking features using random forest.

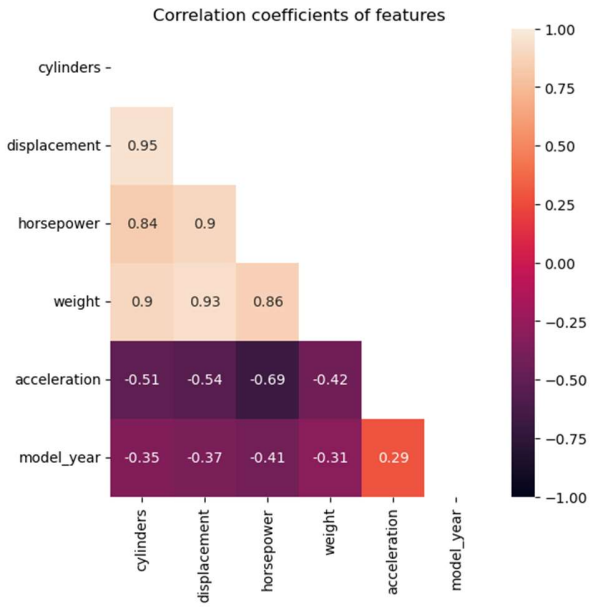

Figure 1: A correlation heat map for features in the Auto MPG dataset.

Methods

Regression

Classification (Logistic Regression)

Logistic regression separates classes using a linear approach to optimise CA. We optimised the hyperparameters using grid search and found the best values for the regularisation parameter (*C*) as 0.01; penalty type as 'L2', solver as 'liblinear'; and maximum iterations as 100.

Regression (Linear Regression)

For linear regression, there were no hyperparameters to tune.

Support Vector Machines

Classification

We have a multi-class problem, so used the 'one-vs-rest' approach as there was no difference in CA compared to 'one-vs-one', and this allows for efficient scalability.

We minimised C while maximising the CA (Figure 2) to produce a more generalisable model but with some trade-off in training accuracy.

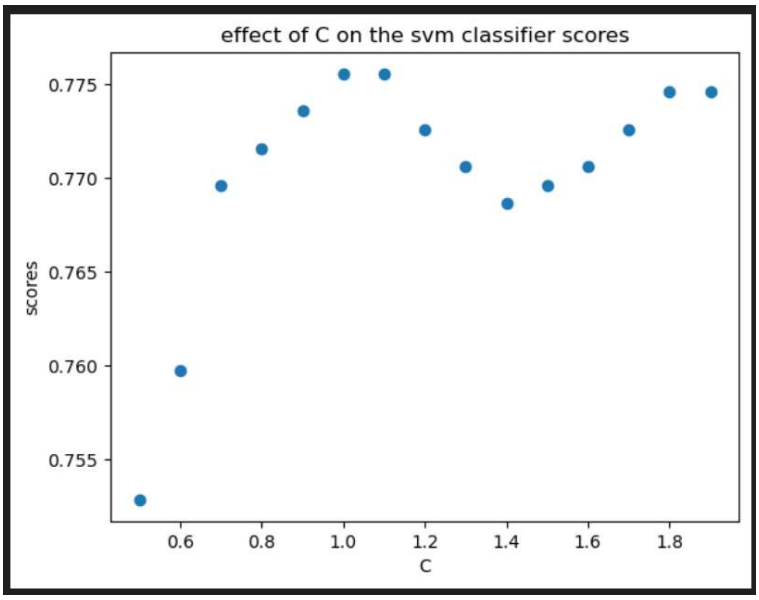

Figure 2: A scatter plot to show the relationship between C and classification accuracy.

Regression

We tuned the hyperparameters ε and C. Figure 3 shows their relationship to the validation score.

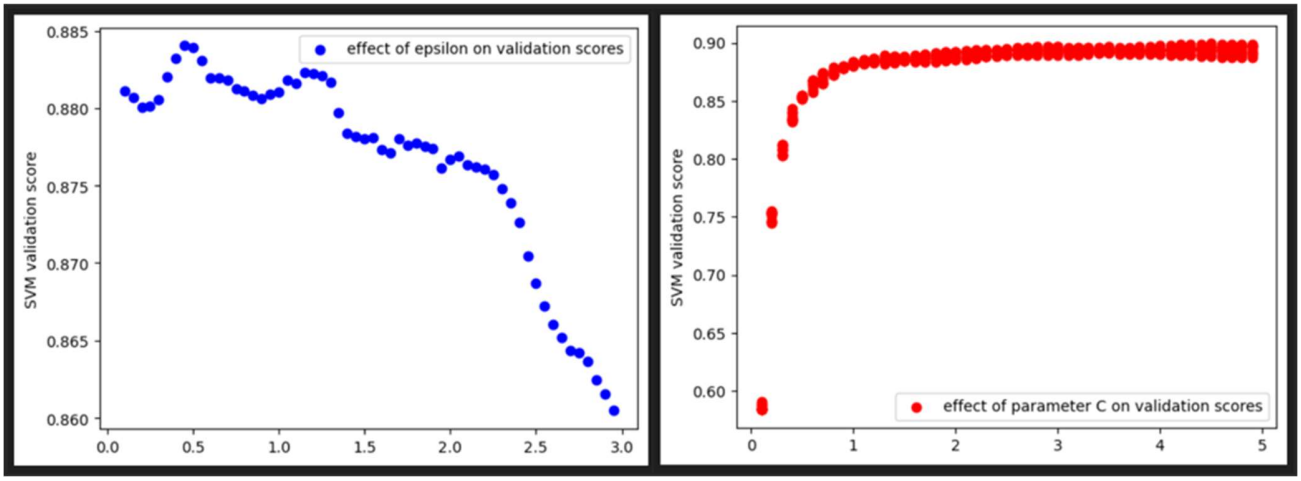

Figure 3: Scatter plots to show the relationship between hyperparameters and model validation score.

We decided to set C=1.0 and ε=0.1 as the improvement in model performance (measured by RMSE) from increasing ε was not significant (<5%) as seen in Table 1.

Table 1: The RMSE of SVR models with different valued hyperparameter epsilon.

	ε = 0.1	ε = 0.5	Error difference	Model improvement /%
RMSE	1.4341	1.4250	0.0091	0.6361

Decision Tree & Random Forest

Classification

For classification, max depths of sixteen for decision tree and twelve for random forest resulted in best accuracy. The number of estimators for random

forest was set to twelve as any larger amount gave comparable results but significantly increased the training time. The confusion matrices for decision tree (Figure 4) and random forest (Figure 5) indicate that both methods have comparable accuracy.

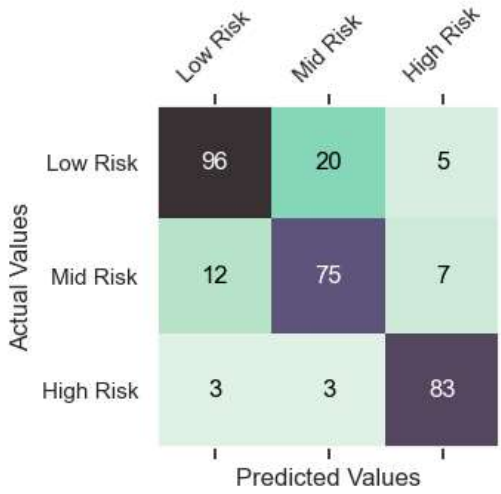

Figure 4: The confusion matrix for the decision tree model.

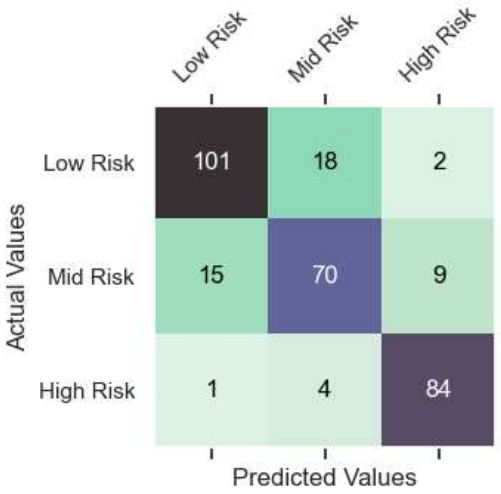

Figure 5: The confusion matrix for the random forest model.

Regression

For regression, max depths of four for decision tree and five for random forest gave the best accuracy, while maintaining low training times. The number of estimators for random forest was set to 100 as any more did not provide better accuracy at the cost of runtime.

Multi-Layer Perceptron Neural Network (MLP)

We used Stochastic Gradient Descent to optimise intrinsic parameters and avoid getting stuck in a local minimum. For classification, a momentum of 0.8 had better accuracy than default at a slight computational cost. For both tasks, the number of iterations was 10,000 to ensure the loss function was minimised.

We tuned hyperparameters through trial-and-error to maximise accuracy and minimise computation time. Setting the regularisation term λ to 1×10^{-4} mitigated overfitting. We chose the ReLU activation function for its efficiency and common use (Nwankpa et al. 2021).

For classification, a constant learning rate (α) was faster however an adaptive α had better cross-validation accuracy scores. For regression, adaptive α was both faster and more accurate.

For regression, two hidden layers with 6 nodes in the first and 4 nodes in the second had best performance. Few layers and nodes were chosen because the data didn't have many dimensions and to avoid overfitting.

Results

Table 2: Results for both tasks, for each model.

	Classification			**Regression**		
	Average training CA (cross-val.)	**CA**	**F1 score**	**Average training MSE (cross val.)**	**Test MSE**	**Test RMSE**
Regression (Linear & Logistic)	0.66	0.65	Low risk: 0.75 Med risk: 0.40 High risk: 0.71	12.15	9.95	3.15
Decision Tree	0.83	0.82	Low risk: 0.82 Med risk: 0.75 High risk: 0.90	14.34	11.47	3.39
Random Forest	0.84	0.84	Low risk: 0.85 Med risk: 0.75 High risk: 0.91	9.12	7.59	2.75
SVM	0.84	0.70	Low risk: 0.76 Med risk: 0.47 High risk: 0.80	10.48	7.09	2.66
ANN	0.78	0.76	Low risk: 0.73 Med risk: 0.67 High risk: 0.88	8.30	6.45	2.54

Conclusion

For classification, random forest had the best accuracy, while logistic regression had the worst. For regression, ANN had the best validation scores, and decision tree had the highest.

References

Ahmed, M. 2020. Maternal Health Risk. UCI Machine Learning Repository. https://doi.org/10.24432/C5DP5D.

Nwankpa, C. E., Ijomah, W., Gachagan, A., & Marshall, S., 2021. Activation functions: comparison of trends in practice and research for deep learning. 124 - 133. 2nd International Conference on Computational Sciences and Technology, Jamshoro, Pakistan.

Quinlan, R. 1993. Auto MPG. UCI Machine Learning Repository. https://doi.org/10.24432/C5859H.