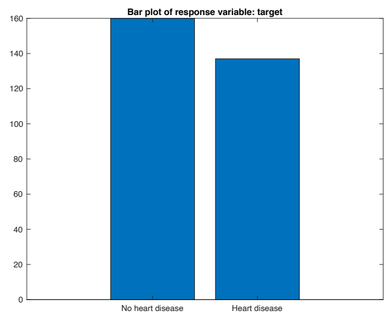
Applying and comparing Naïve bayes (NB) and Logistic regression (LR) on the heart disease UCI dataset

Description and motivation of the task:

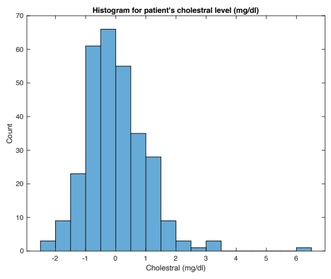
I will aim to perform a binary classification task in order to predict whether a patient has heart disease or not based on the features in the dataset, by applying two models using machine learning algorithms: Naïve bayes and Logistic regression. Labels are given in the dataset which makes the task a supervised learning task. I will then compare the two machine learning models to see which model performed better using performance metrics such as accuracy. I will also compare results obtained by (Rawat, 2019) [1].

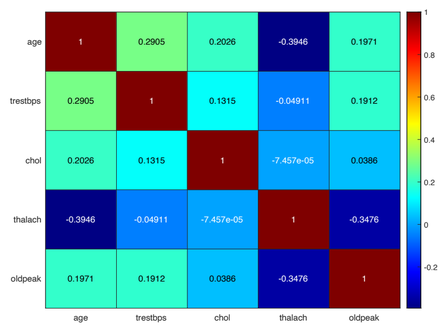
Analysis of the dataset:

- Cleveland dataset contains information about patients from a health clinic in Cleveland which was downloaded from the UCI Machine learning repository.
- The dataset originally contained 303 rows and 14 columns, I removed 6 observations due to missing datapoints. The Final cleaned dataset contains 297 observations (rows) and 14 features (columns): which has 5 numeric features, 8 categorical (nominal) features and a categorical (ordinal) feature.
- I derived a new column ”target” from the original 14th variable “num” in order to covert the multiclass values in the ”num” variable to binary values. The multiclass values are defined by the presence of heart disease in order: no presence = 0 and having presence of heart disease = 1,2,3,4 . I replaced values 2,3 and 4 with 1 to convert into a binary values : No heart disease = 0 and Heart disease = 1. This was placed in a new column “target” replacing the “num” column as the task becomes a binary classification task due to this manipulation.
- I acknowledged that I have noisy data in the form of outliers which was detected by boxplots.
- The dataset has a slight imbalanced class problem as the class of the target variable is shown by the bar plot.
- Most of the numeric features follow normal distribution this was seen via the histograms; this was more notable when I standardized the numeric features for the dataset.
- All numeric feature do not have strong correlation (<0.50) as seen by the heatmap which suggest strong independence which is good for my algorithms. There is no case to drop any features from the dataset .



	Mean		Standard deviation		Range		IQR	
	Heart disease (1)	No heart disease (0)	Heart disease (1)	No heart disease (0)	Heart disease (1)	No heart disease (0)	Heart disease (1)	No heart disease (0)
age (years)	56.76	52.64	7.9	9.55	42	47	9.25	14.5
trestbps (mmHG)	134.64	129.18	18.9	16.37	100	86	25	20
chol (mg/dl)	251.85	243.49	49.68	53.76	278	438	66.75	60
thalach (BPM)	139.11	158.58	22.71	19.04	124	106	32.25	23
oldpeak	1.59	0.6	1.31	0.79	6.2	4.2	1.95	1.1





Logistic Regression:

- Logistic regression is a supervised learning classification model used to classify binary values to predict whether a patient has class: heart disease = 1 or no heart disease = 0.
- Defining a class to be positive or negative is arbitrary and is our choice to make based on interpretation of the class.
- Logistic regression is a type of Generalized linear model (GLM)
- The class is determined by the posterior probability calculated from the logit function. Where the threshold my case is 0.5 (50%): if the posterior probability is greater than 0.5 then the class will be 1 (heart disease), otherwise the class will be 0 (no heart disease).
- Logistic regression can use lasso, Ridge or Elastic Net regularization to prevent the model from overfitting and to reduce complexity of the model by feature engineering or handling multicollinearity. This helps as it makes “The model work harder to explain the training data” (Pocs, 2016) [3] , so we can get a better fit and generalization on the test data.
- Logistic regression could use maximum likelihood estimator to get a best fit for the model and estimate the parameters.

Pros

1. Can extend to multiclass classification task including ordinal class data.
2. Can be transformed from linear models using the sigmoid/logistic function, as it is related to linear regression.
3. Can implement both continuous and discrete variables in the logistic regression model.
4. Does not require any scaling for the data.
5. Quick to train and to classify as the results are interpretable such that you can see measure of a predictor by coefficient size as well as if the impact is positive or negative, the coefficient also show indicators of feature importance.
6. exponentiated logistic regression slope coefficient can be interpreted as an odds ratio, which can indicate how much odds of an outcome can change for a unit increase in the predictor feature (Schober and Vetter, 2021) [4].
7. Makes no assumption about distribution of classes in the feature space.

Cons:

1. Limitation is that it has assumption of linearity between dependent & independent variables.
2. Non-linear problems cannot be solved due to linear decision boundary.
3. It needs the independent variables to be linearly related to log odd (log(p/1-p)).
4. Can be Affected by outliers and multicollinearity and missing values

Naïve Bayes:

- Naïve Bayes is another supervised learning classification model used to classify a feature. It has a ‘Naïve’ assumption that the feature variables used to classify the response variable are independent.
- Naïve bayes uses bayes theorem to calculate the posterior probability, this uses the likelihood probability, class prior probability. The posterior probability is then used to predict the class.
- Naïve bayes is the simplest form of the Bayesian network (Dhotre and Kaviani, 2017) [9]
- Naïve bayes works under so many distributions such as multinomial and Bernoulli distribution for continuous variables, for categorical/discrete variables it assumes a multinomial distribution.
- Prior calculated from the response variable which is probability of it being from a certain class.

Pros

1. Easy to interpret, use and efficient to predict a class.
2. Can be used on a small dataset as it can be trained on small training set than other machine learning algorithms.
3. Despite the naïve assumption, the algorithm preforms well.
4. Handles high dimensional data well.
5. Works with many distributions.

Cons:

1. Works better when you perform feature scaling.
2. The assumption of independence between predictor features rarely holds, especially for big datasets. This is where other complex algorithms outperforms naïve bayes.
3. If the probability for the likelihood equals 0, meaning algorithm fails to make a prediction. To work around this, you would need to apply a smoothing technique
4. Work better with categorical/discrete variables rather than continuous.

Hypothesis:

- ❑ I assume that naïve bayes will outperform logistic regression since the dataset is small with 297 obversions and 13 features used to predict the response feature. This is coupled with the fact that logistic regression be affected by outliers and missing values, which we have detailed in the preprocessing steps.
- ❑ I expect both models to do well as the heatmap shows that no predictor features have any strong correlation, indicating that the naïve assumption of independence for predictor features holds, and also because (Rawat, 2019) and Amin, M.S., Chiam, Y.K. & Varathan, K.D. 2019 [2] got reasonably good results for the respective algorithms.

Methodology:

1. Split the dataset: 80% for the training set and 20% for the test set just like (Rawat, 2019) [1] .
2. Apply, cross validation on the training data by applying K-folds, where K=10. To get an expected fit of the model through calculating the K-Fold Loss.
3. Fit the model using the training data, then predict the classification label using the test data.
4. Optimize the model using hyperparameter tuning to sub the optimized parameters to make the models optimal.
5. Determine whether the model is optimal by comparing it to the baseline model through seeing if the accuracy increased while seeing if the training error and cross-validation error decreased.
6. Assess the performance models by performance metrics such as accuracy , ROC Curves, AUC and confusion charts and compare the results with (Rawat, 2019) [1] .

Choice of parameters and experimental results:

For both Naïve bayes and Logistic regressions models I selected all 13 features and did no feature selections. It also has a slight imbalance classification problem, but due to it being slightly imbalance, I didn’t have to rebalance the classification of outcomes.

AUC has a value between 1 and 0 with 1 being the perfect performance and 0 being the worst.

Logistic regression

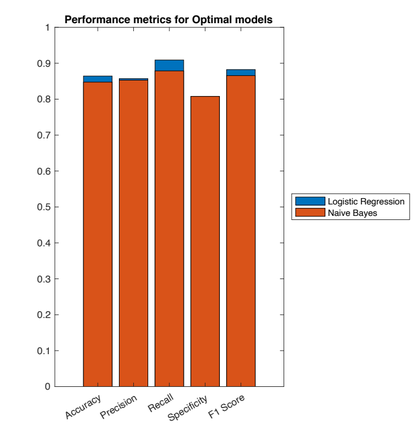
- Parameters chosen was the lambda value and ridge regression regularization type under hyperparameter tuning to optimize the model. This helped optimize the results from the baseline model by fitting the model through penalizing the degree of the function.

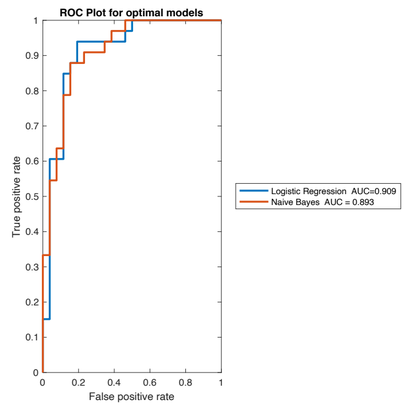
Naïve Bayes

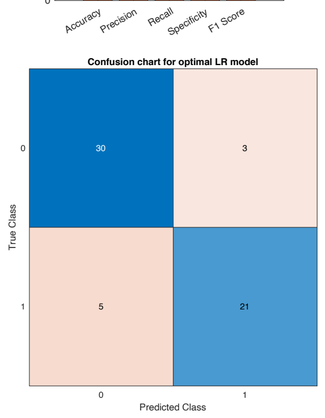
- From the baseline model I chose to fit the distribution with the Gaussian (normal) distribution.
- In the optimal model, the choice of parameters was distribution name and width size, where the distribution changed from normal distribution to kernel distribution and a width size, where scaling was needed to get an accurate value for the width size.

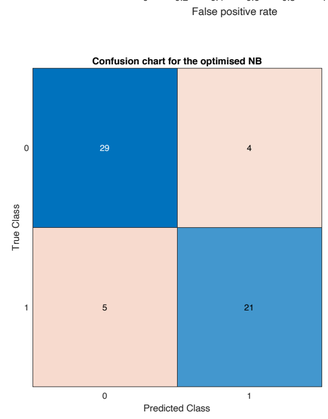
	LR Baseline	NB Baseline
Cross validation error	0.1639	0.1723
Training error	0.1218	0.1513
AUC value	0.909	0.893

	Optimal LR	Optimal NB
Accuracy	0.8644	0.8475
F1-Score	0.8824	0.8788
Precision	0.8571	0.8529
Recall	0.9091	0.8788
Cross validation error	0.1429	0.1597
Training error	0.1261	0.1513
Test error	0.1356	0.1525









Analysis and critical evaluation of results:

- ➢ For the baseline models, logistic regression has a lower cross validation error than naïve bayes meaning that it has a better expected fit by 0.84% (0.1723 – 0.1639). Where the cross-validation error for LR was 16.39% and for NB 17.23%. This show that both models are expected to perform well in relation to one another, with LR expected to perform marginally better.
- ➢ This was the case for the baseline model as the training error for LR was at 12.18% while NB has a training error of 15.13%, as the AUC value for LR was 0.909 performing better than NB which has an AUC value of 0.893 for the test set.
- ➢ NB may have had a higher training error due to the naïve assumption of independence between predictor features as some variables has a hint of correlation. To add to this, it could be that LR can handle the complexity of the model as all 13 features was selected.
- ➢ One of the ways to see if my models had optimized was seeing if my cross-validation error and training error minimized. For NB, the training error stayed the same, but my cross-validation error reduced from 17.23% to 15.97% showing an improved fit. This also occurred for LR as the cross-validation error went from 16.39% to 14.29%, In fact it has improved fit (2.1%) than NB (1.26%). However, LR training error increased by 0.43% (0.1261 – 01218), this could be due to the model fitting the training data being affected by the noise of the data and this will cause high bias/variance for LR as NB would be unaffected by the noise in the dataset. Either way, both models have been optimized as they have a better expected fit through cross validation.
- ➢ The optimal LR model performed better on the test set than NB, with LR having an accuracy of 86.44% whilst NB has an accuracy of 84.75%, this was also confirmed through the loss (error) where LR has an error of 13.56% as NB has an error of 15.25%. This shows that both my models performed well and were very close in terms of accuracy however LR still did perform better than NB, which was not expected (hypothesis statement) , this could be perhaps due to the complexity of the models , with using all 13 features for both models as LR would be able to handle the complexity of the model than naïve bayes, as the more features there are the less likely the “naïve” assumption of independence will hold . I could further optimize the models by feature selection in order to reduce the complexity of the model and therefore have a better fit for the models.
- ➢ The performance metrics such as accuracy was calculated using the TP,TN,FP,FN cases instead of the in-built MATLAB function. This is so I can make a direct comparison with the class cases being positive or negative with the performance metrics.
- ➢ When comparing the accuracies on the test set for the models with (Rawat, 2019) , where the accuracy for LR on test set = 80.32% and for NB accuracy on test set = 78.68%, This fellow publisher also had LR performing marginally better than NB , however I had a larger accuracy for LR (86.44%) and NB (84.75%) this could be due to my models being optimized as well as the random train/test split.
- ➢ I used the TP,TN,FP,FN to calculate the performance metrics accuracy, precision, recall, specificity and the F1 score and to also verify that my confusion chart was correct. I defined the positive case to be of class = 0 (No heart disease) & negative case to be of class = 1 (Heart disease), this is arbitrary and open to interpretation. The precision value is higher for LR (85.71%) than for NB (85.29%) showing that LR predicting the positive case of a person not having heart disease (the positive case) so if the clinic wants to priorities predicting patients having no heart disease over predicting they have heart disease, they should pick LR.
- ➢ Recall, has the higher value for (90.91%) LR than NB (87.88%) , it’s also shown by the bar plot. This shows that LR is better at predicting the positive class. F1 score for LR is just higher than NB as shown by bar plot, which also measures accuracy but for more balanced classification set as our response variable was slightly imbalanced.
- ➢ The AUC value (accurate to 3 decimal place) remains the same from the baseline for NB and also for LR. The AUC is calculated from the roc curve which shows that LR curve performs than NB due to the lines showing it has better trade off at the TPR and FPR.

Lessons learned:

- I learnt that model performance is not just based of on accuracy as it is on other factors such as F1 score, recall and precision in case you want to priorities predicting the positive case over the negative case , for instance the clinic may want to predict patients with heart disease more correctly than people that do not, so they can treat more patients. Model performance is dependent of the goal of the user.
- I also learnt that model building is an iterative process in order to maximize your accuracy to predict an event occurring as this depends on the fit of your model through balancing bias and variance.

Future work:

- I could optimize my models further through feature selection to reduce the complexity of my model and gain a better fit for the model, this can be done using the steps outlined by Amin, M.S., Chiam, Y.K. & Varathan, K.D. 2019 [2].
- I could also merge other datasets from other clinics in different locations provided in the UCI machine learning repository ,with myself selecting a dataset from a singular clinic. This would increase the number of observations allowing myself to split the dataset into a training set, validation set, and test set rather than just a train and test set. The validation set will allow me to validate my models further.

References :

[1] (Rawat, 2019) Rawat, S., 2019. *Heart Disease Prediction*. [online] Medium. Available at: <https://towardsdatascience.com/heart-disease-prediction-73468d630cfc> [Accessed 15 December 2021].

[2] Amin, M.S., Chiam, Y.K. & Varathan, K.D. 2019, "Identification of significant features and data mining techniques in predicting heart disease", *Telematics and informatics,* vol. 36, pp. 82-93.

[3] (Pocs, 2016) Pocs, M., 2016. *Hyperparameter Tuning in Lasso and Ridge Regressions*. [online] Medium. Available at: <https://towardsdatascience.com/hyperparameter-tuning-in-lasso-and-ridge-regressions-70a4b158ae6d> [Accessed 15 December 2021].

[4] (Schober and Vetter, 2021) Schober, P. and Vetter, T., 2021. Logistic Regression in Medical Research. *Anesthesia & Analgesia*, 132(2), pp.365-366.

[5] (Peng, Lee and Ingersoll, 2002) Peng, C., Lee, K. and Ingersoll, G., 2002. An Introduction to Logistic Regression Analysis and Reporting. *The Journal of Educational Research*, 96(1), pp.3-14.

[6] (RanjanRout, 2021) RanjanRout, A., 2021. *Advantages and Disadvantages of Logistic Regression - GeeksforGeeks*. [online] GeeksforGeeks. Available at: <https://www.geeksforgeeks.org/advantages-and-disadvantages-of-logistic-regression/> [Accessed 15 December 2021].

[7] Murphy. K, “ *Machine Learning A Probabilistic Perspective “*, 2012.

[8] (Zhang, 2021) Zhang, Z., 2021. *Naive Bayes Explained*. [online] Medium. Available at: <https://towardsdatascience.com/naive-bayes-explained-9d2b96f4a9c0> [Accessed 15 December 2021].

[9] (Dhotre and Kaviani, 2017) Dhotre, S. and Kaviani, P., 2017. SHORT SURVEY ON NAIVE BAYES ALGORITHM. *International Journal of Advance Engineering and Research Development*, [online] 4(11). Available at: <https://www.researchgate.net/publication/323946641_Short_Survey_on_Naive_Bayes_Algorithm>.

[Dataset] https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/