Applied Bayesian Statistics

WINE QUALITY PREDICTION USING BAYESIAN METHODS

Vu Kieu Chinh

RMIT University,

School of Science

2024

Executive Summary

This report utlises the Bayes' methods to induce classifications for qualities of white wines, using various prior distributions. A total of approximately 5,000 samples of white wines were used and split for model training and the test of the classification model. Two different prior distributions, non-informative and informative, were used to obtain the posterior distributions. Furthermore, different MCMC settings for chain generations were implemented to improve efficiency. The MCMC diagnostic tests were performed on the betas, zbetas, and predictions to ensure that the generated chains were representative and accurate. As the diagnostic plots illustrated representativeness and accuracy of chains, the posterior distributions were used for hypothesis tests and obtain Bayes' intervals and point estimates.

The sensitivity analyses were performed to determine the robustness of prior distributions' specifications. The classification results showed that both the non-informative prior and informative prior produced rather excellent results. They had approximately similar results of 78% of accuracy, 86% of precision, 86% of recall and an F-score of 86%. Thus, the findings of this research was successfully achieved with the classifications of white wines and highlighted specific properties that wine producers could use to enhance the quality.

Table of Contents

I. Introduction

The Bayesian methods could be utilised in numerous different ways such as making predictions or regression analyses on numerous subjects of interest. Classification is another practical utilization of the Bayesian method as it allows for incorporation prior knowledge of the subject of interests with a degree of belief.

The purpose of this report is to make predictions on the classifications of quality for white wines, using numerous independent variables. The classification is done using the Bayesian's methods by obtaining posterior distributions with different prior distributions and a sample data, the likelihood. This report will discuss the successful outcomes of the classification, including the accuracy rate, precision and recall rate, and comparisons of classification results will be made for non-informative and informative prior distributions. The aim for the classification is to identify two distinct types of white wines, "High quality" and "Not high quality".

II. Methodology & Data

The 'Just Another Gibbs Sampling' (JAGS) program and RStudio will be specifically utilised to make predictions of the classification of the wine quality.

The sample data used for this report was retrieved from the "Modeling wine preferences by data mining from physicochemical properties" research by Cortez et al. (2009). The sample contains approximately 5,000 observations and 12 variables in total. 11 of the variables within the data set will be treated as independent variables, while the 'Rating' variable will be treated as the dependent variables. These independent variables include fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol. All the independent variables from the sample data will be included in the classification model as they are relevant to the characteristics of white wines.

The 'Rating' dependent variable is required to be modified, as each white wine was assigned a rating score from 1 to 10. This means that this dependent variable is an ordinal data type. However, the classification model for the white wine would require binary values to denote "High quality" and "Not high quality". Thus, the modifications will be made as follows:

- White wines that received ratings of 7 and above will be assigned a value of 1, and
- White wines that received ratings of 6 and below will be assigned a value of 0.

The values of 0, and 1 were given to denote the quality of the white wines, i.e. "high quality" and "not high quality" respectively.

III. Discussion

1. Descriptive Look

This section aims to explore and discuss the nature of the wine quality contained in the sample data.

Expectation of classification results: In total, the sample data consists of 1,060 “High quality” white wines and 3,838 of “Not high quality” white wines, or approximately 22% and 88%, respectively. Thus, it is expected that the predicted classifications of the white wines will follow this proportion of the sample data. In other words, in the best-case scenario, the JAGS model will be able to classify the quality of white wine with 22% “High quality” and 88% “Not high quality”.

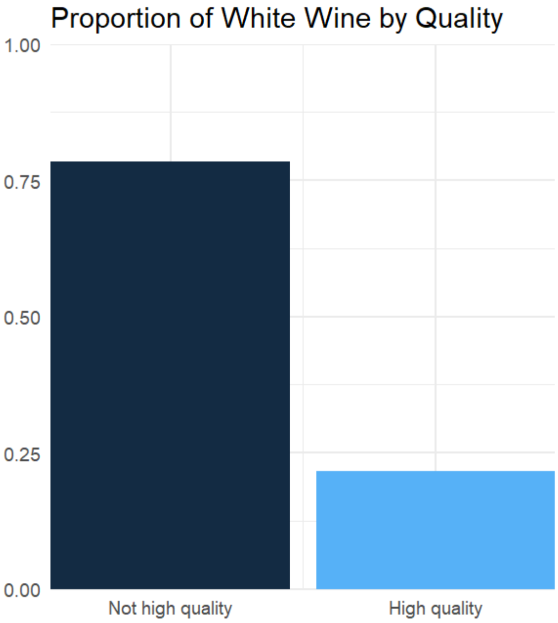
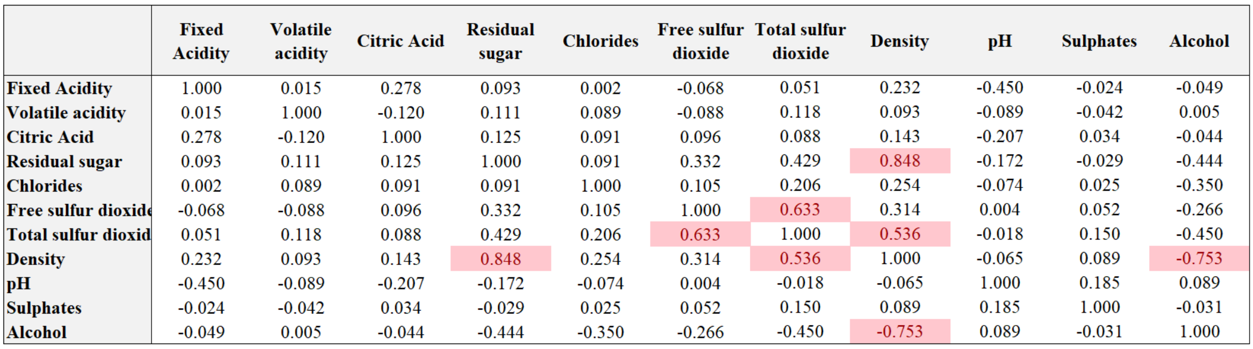


Figure 1: Proportion of white wine by quality from the sample

Correlation matrix of independent variables

Table 1: Correlation matrix of independent variables (strong correlation shown in red)



	Fixed Acidity	Volatile acidity	Citric Acid	Residual sugar	Chlorides	Free sulfur dioxide	Total sulfur dioxide	Density	pH	Sulphates	Alcohol
Fixed Acidity	1.000	0.015	0.278	0.093	0.002	-0.068	0.051	0.232	-0.450	-0.024	-0.049
Volatile acidity	0.015	1.000	-0.120	0.111	0.089	-0.088	0.118	0.093	-0.089	-0.042	0.005
Citric Acid	0.278	-0.120	1.000	0.125	0.091	0.096	0.088	0.143	-0.207	0.034	-0.044
Residual sugar	0.093	0.111	0.125	1.000	0.091	0.332	0.429	0.848	-0.172	-0.029	-0.444
Chlorides	0.002	0.089	0.091	0.091	1.000	0.105	0.206	0.254	-0.074	0.025	-0.350
Free sulfur dioxide	-0.068	-0.088	0.096	0.332	0.105	1.000	0.633	0.314	0.004	0.052	-0.266
Total sulfur dioxid	0.051	0.118	0.088	0.429	0.206	0.633	1.000	0.536	-0.018	0.150	-0.450
Density	0.232	0.093	0.143	0.848	0.254	0.314	0.536	1.000	-0.065	0.089	-0.753
pH	-0.450	-0.089	-0.207	-0.172	-0.074	0.004	-0.018	-0.065	1.000	0.185	0.089
Sulphates	-0.024	-0.042	0.034	-0.029	0.025	0.052	0.150	0.089	0.185	1.000	-0.031
Alcohol	-0.049	0.005	-0.044	-0.444	-0.350	-0.266	-0.450	-0.753	0.089	-0.031	1.000

Table 1 shows correlations between independent variables used for this report. Most of the correlations between these predictors are low. However, there are several notably strong positive and negative correlations shown in red. These correlations will have adverse impacts on the representativeness and accuracy of chains. Thus, standardisations methods and other techniques to reduce autocorrelation, such as the use of further thinning steps, will be implemented to reduce the adverse impacts of correlations between predictor variables.

Independent variables versus quality of wines:

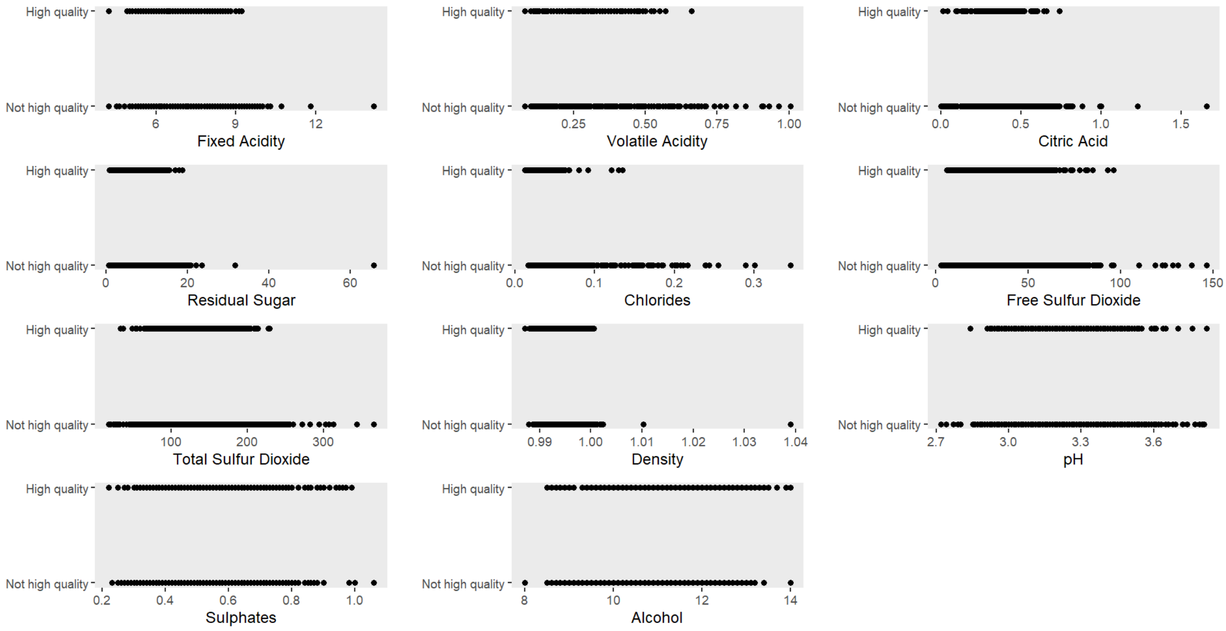


Figure 2: Descriptive look of independent variables and white wine quality

Figure 2 illustrates the difference in characteristics of the independent variables for the "High quality" and "Not high quality" white wines. Although, for some independent variables, the two classifications of wine cannot be directly differentiated visually using the scatter plots above, there are other notable differences. For instance, some "High quality" wines tend to have lower volatile acid level, while some "Not high quality" wines have volatile acid level up to 1 (unit). Interestingly, independent variables such as Chlorides, Free sulfur dioxide and Total sulfur dioxide follow the same pattern as the Volatile acid level. In contrast, there are no notable differences between the two qualities of wines for independent variables such as Residual sugars, Density, Sulphates, and Alcohol levels. Thus, it is expected that these variables will have minor impact on the classification model.

Outliers: Upon the inspection of the scatter plot above, there are several noticeable outliers within the data set that could have significant impact on the prediction of the model. Specifically, independent variables such as Citric acid, Residual sugars, Chlorides and Density have significant outliers that should be addressed. Figure 3 illustrates scatter plots that highlight the outliers, shown in red, within some of the independent variables. These outliers will have some adverse impacts on the training of the model, which will reduce its capability of making correct predictions/classifications.

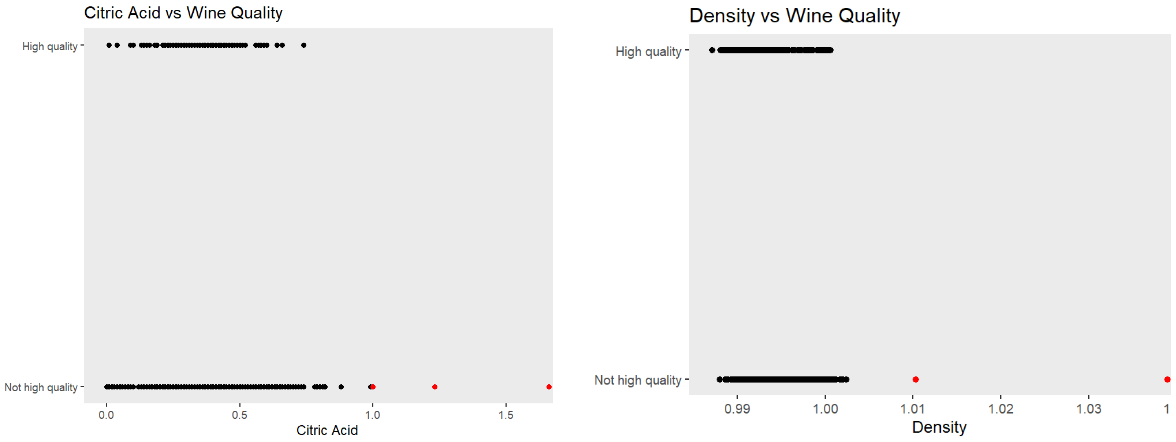


Figure 3: Potential outliers (shown in red) within Citric acid and Density independent variable

Guess parameters are incorporated into the model to reduce the impact of outliers on the model and MCMC classification. The specifications of the guess parameters shall be thoroughly discussed in the JAGS models section of this report.

Training and Prediction Data: The sample data is split into two separate files. The purpose of splitting the sample data is to train the model, while retaining a smaller portion of the sample that the model will make predictions for. The proportion of the split will be as follows:

- 70% of the sample data is used to train the model, and
- 30% remaining of the sample data will be used to make predictions.

This split proportion is chosen to ensure that the classification results will be as accurate and representative as possible. This is because model training usually requires many observations and large iterations of chain generations. Thus, the larger split is assigned to the training of the model. However, it is expected that the remaining 30% of the sample data will include some characteristics of the independent variables that are not available in the model-training split. Thus, the classification results may not be 100% correct but it is acceptable if the model is able to make a substantial portion of correct classifications.

2. Mathematical Model & JAGS Model Diagram

In Figure 4 diagram, the observed data point Y_i, representing the classification of wine quality, follows a Bernoulli distribution. The dependent variable is binary, taking value 0 to indicate "Not high quality" and 1 to indicate "High quality". To model the probability of a "High quality" classification, μ of the Bernoulli distribution is defined by the logistic function. The logistic regression achieves this by combining the predictors x_j with their corresponding coefficients β_i, along with an intercept β_0, forming the expression:

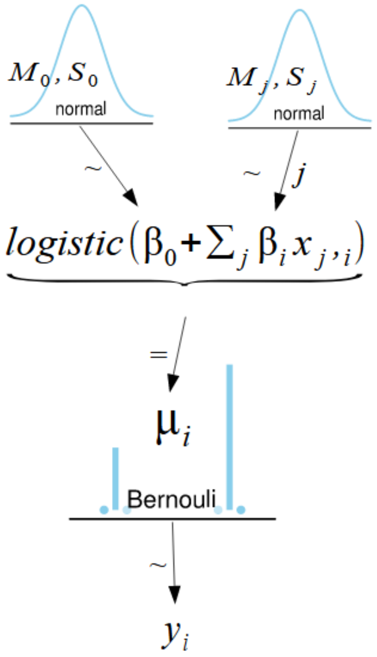


Figure 4: JAGS model diagram for the logistic regression model

Y = logistics (β_0 + β_1 Fixed acidity + β_2 Volatile acidity + β_3 citric acid + β_4 residual sugar + β_5 chlorides + β_6 free sulfur dioxide+ β_7 total sulfur dioxide+ β_8 density+ β_9 pH+ β_{10} sulphates+ β_{11} alcohol)

The coefficients β_i represent changes in the log-odds of the dependent variable being classified as "high quality" versus "not high quality". For each unit increase in an independent variable, the corresponding β_i value informs how much the log-odds increase or decrease. The odds ratio, $\exp(\beta_i)$, represents the multiplicative change in odds for a unit change in the predictor.

In applying Bayesian logistic regression, this study assigns prior distributions to the model parameters. Each regression coefficient has a normal prior distribution with mean M_0 for the intercept and M_j for each predictor, along with standard deviations S_0, S_j. These normal priors are ideal because they allow the coefficients to range from -∞ to +∞, capturing both positive and negative effects on the outcome Y_i.

Furthermore, as discussed earlier, the guess parameter is incorporated to mitigate the influence of outliers in the dataset. In this approach, the data is viewed as a blend of two distinct sources. One source arises from the logistic function applied to the predictors, while the other represents randomness or "guessing". We add a new parameter α, which denotes the probability that a data point results from the guessing process. Consequently, with a probability of 1 - α, the Y value is generated by the logistic regression model. The model is then constructed by merging this random guessing component with the logistic regression, as shown below:

$$\mu = \alpha \text{ x } 1/2 + (1 - \alpha) \text{ x logistic } \left(\beta_0 + \sum_{i=1}^{k} \beta_i\right)$$

3. Specify The Prior Distribution

In this Bayesian logistic regression model, the dependent variable is binary, representing the quality classification of white wine.

The independent variables (or predictors) are continuous, representing the physicochemical properties of the wine, which are used to predict the wine's quality.

3.1 Non-informative prior

For non-informative prior distributions, expert's information/knowledge and their degree of beliefs are not incorporated into the model. The non-informative priors are applied to each coefficient and the intercept, assuming normal distributions with a mean of 0 and large variance, which reflects the absence of strong prior beliefs about parameters.

=> $\beta_0 \sim N\ (0, 2)$

$\beta_i \sim N\ (0, 2)$

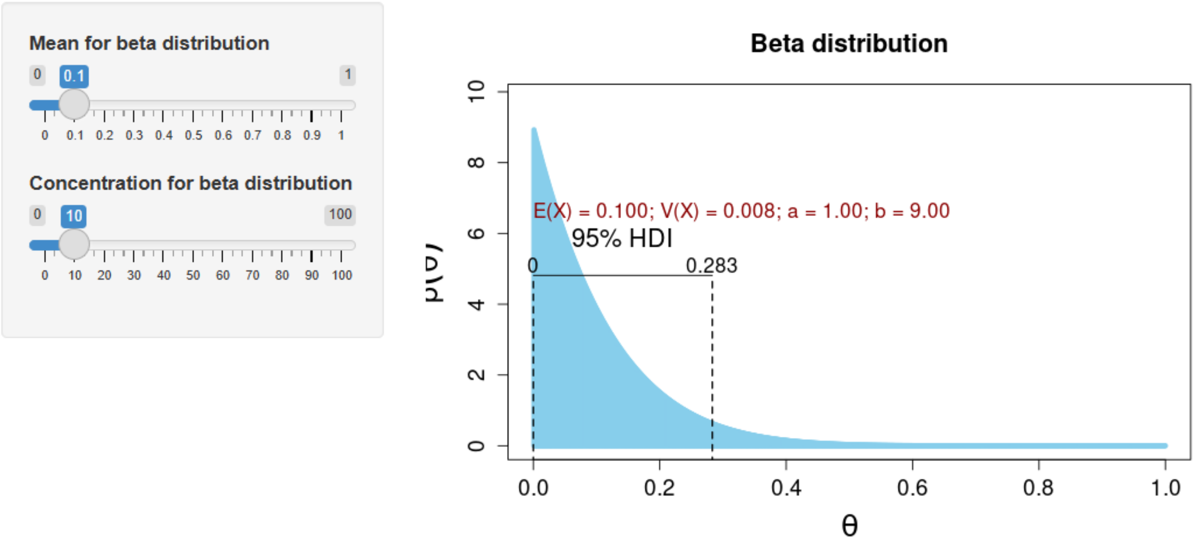
The mean of 0 is chosen to ensure that the prior distribution will have no effect on the posterior distributions, while the large variance is chosen to ensure to the posterior is not concentrated around this given mean value.

Expectation of the posterior distribution using non-informative prior distributions: it is expected that the posterior distributions will be dominated by the likelihood, or the sample data. This is because the non-informative prior distributions will have no influence on the posterior distribution, given their large variances around the means of 0.

A **sensitivity analysis** on the non-informative prior distributions will be performed and discussed later in this report. The purpose of the sensitivity analysis is to determine the validity of the specifications of the prior distribution above.

Guess parameters for non-informative prior: as discussed earlier, outliers within the data set will be dealt with using guess parameters. It follows a Beta distribution with alpha set to 1 and beta set to 9, which keeps guess predominantly low but allows flexibility, enabling the model to adjust for deviations in the data.

=> guess ~ dbeta (1, 9)



Mean for beta distribution

0.1

Concentration for beta distribution

10

Figure 5: Beta distribution for guess parameter with non-informative priors

3.2 Informative prior

In constructing the Bayesian model, this study incorporated prior knowledge about feature importance from the research “Modeling wine preferences by data mining from physicochemical properties” by Cortez et al. (2009), and the direction of effects into the prior distributions of the coefficients. Using the logistic regression results as a reference, each predictor's prior mean was assigned based on its feature importance, scaled relative to the most important feature, such as sulphates, while maintaining the positive or negative sign of each effect as indicated by the regression results. For instance, sulphates and alcohol, which showed high importance and positive coefficients, were given higher positive prior means, whereas volatile acidity and chlorides, with strong negative coefficients, were assigned scaled negative means.

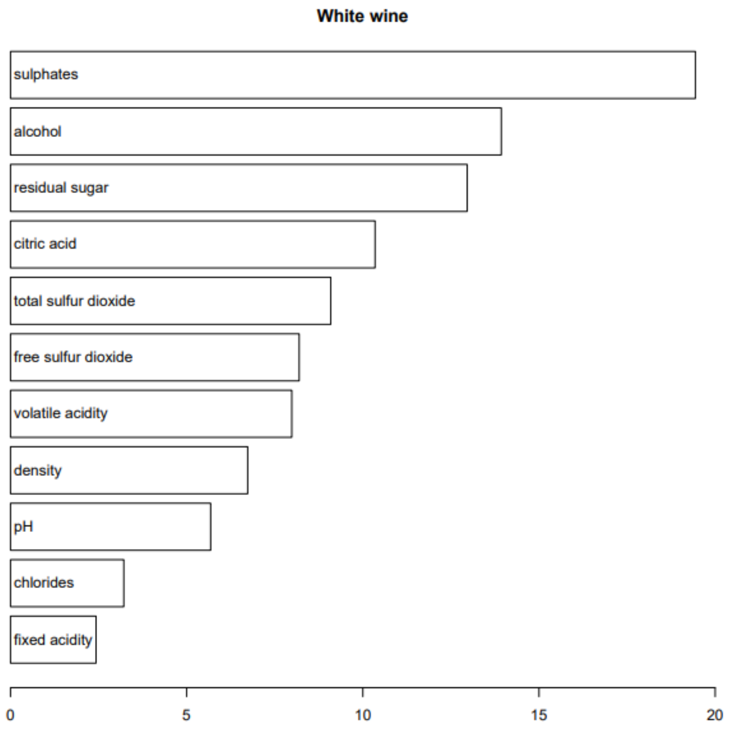


Figure 6: White wine input importances (in %)

To reflect the importance, the prior means for each feature are scaled accordingly. Features with higher importance, like sulphates and alcohol, received larger prior means, while less important features, like fixed acidity, had their means scaled down. For example, sulphates, the most influential feature with 20% importance, had its prior mean doubled, while other features were scaled proportionally up (multiplied by a value greater than 1) or down (multiplied by a value less than 1) based on their relative importance. This approach ensures that the model starts with informed expectations, allowing more significant features to exert a greater influence on the outcome, reflecting their relative impact. The variance for these priors was set relatively high at 1 to account for the uncertainty in the prior information. Further reasoning for this choice will be explained in the sensitivity analysis that follows.

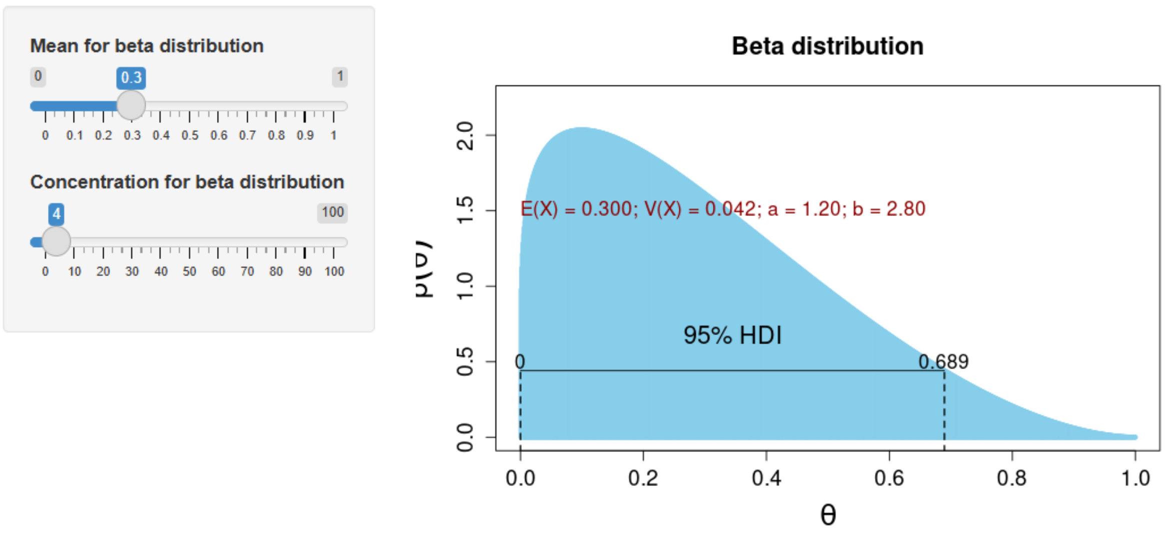
Table 2: Informative priors based on feature importance for Bayesian logistic regression model

Feature	**Coefficient**	**Importance (%)**	**Scaled Prior Means**
Fixed acidity (β_1)	0.484	2	0.0968
Volatile acidity (β_2)	-2.944	8	-2.3552
Citric acid (β_3)	-0.795	11	-0.8745
Residual sugar (β_4)	0.304	13	0.3952
Chlorides (β_5)	-7.787	3	-2.3361
Free sulfur dioxide (β_6)	0.0105	8	0.0084

Total sulfur dioxide (β7)	-0.00146	9	-0.001314
Density (β8)	-683	7	-478.1
PH (β9)	3.599	6	2.1594
Sulphates (β10)	2.685	20	5.37
Alcohol (β11)	0.204	14	0.2856

Guess parameter for informative prior: In the informative case, a different guess parameter configuration is applied to address a higher presence of outliers by increasing the means of the distribution. Here, alpha is set to 1.2, and beta is set to 2.8.

=> guess ~ dbeta (1.2, 2.8)



Mean for beta distribution 0.3

Concentration for beta distribution 4

Figure 7: Beta distribution for guess parameter with informative priors

4. Find The Posterior Distribution

4.1 MCMC Settings and Improvements to Achieve Efficiency

As different MCMC settings will directly affect the representativeness, accuracy and efficiency of chain generations, several combinations of MCMC settings were used for both non-informative prior and informative prior. However, it is essential to obtain an efficient setting to generate MCMC chains, without sacrificing representativeness and accuracy of the posterior distribution.

Large settings of 5,000 burn-in steps and 10,000 steps were used to ensure the convergence of chains. The large adaptation steps allowed JAGS to be well-trained to the model and produced representative posterior distributions. On the other hand, the large burn-in steps allowed JAGS to drop the start of the chains, where they had not yet converged. Furthermore, a large thinningstep of 50 was opted for to reduce the autocorrelations between the independent variables. This is done to ensure accuracy of chain and minimising the shrink factors to be lower than 1.2.

The followings are the summary of the MCMC settings:

- Number of burn-in steps:	= 5,000
- Number of adaptation steps:	= 10,000
- Number of thinning steps:	= 50
- Number of chains:	= 2
- Number of saved steps:	= 5,000
- Total number of iterations:	= 250,000

Improvements to settings and further thinning steps: another MCMC settings were tested and used to determine whether more efficiency could be achieved, while retaining similar representativeness and efficiency. It is noted that the large saved-step setting hindered efficiency, and thus, was reduced to 3,000. This resulted in a smaller total of iterations of 150,000. However, the number of burn-in steps and adaptation steps were kept constant as the settings from above. This is done to ensure that the model is well-trained and convergence of chains. The following is the summary of the smaller MCMC settings:

- Number of burn-in steps:	= 5,000
- Number of adaptation steps:	= 10,000
- Number of thinning steps:	= 50
- Number of chains:	= 2
- Number of saved steps:	= 3,000
- Total number of iterations:	= 150,000
- *Further thinning steps:*	*= 3*

Further thinning: these smaller settings produced similar results to the initial settings, while being reasonably more efficient due to the smaller number of total iterations. However, some autocorrelations persisted, and thus, further thinning steps of 3 were used. The further thinning steps had reduced to autocorrelations and improved the shrink factor, while maintaining convergence of chains resulting in acceptable representativeness and accuracy while being more efficient.

Thus, the diagnostics and posterior distributions for the non-informative prior distribution will be based on the smaller settings with the use of thinning steps. In contrast, the diagnostics and posterior distribution for the informative will be based on the improved settings, without the farthing thinning steps. Details of the results will be discussed in the following section.

4.2 Non-informative

4.2.1 Diagnostic tests for non-informative priors

This section aims to conduct the diagnostic tests of the MCMC prior to obtaining the posterior distributions and drawing inferences on the Bayes' point estimates and point interval. The appropriateness of the chains generated for the parameters will be determined to ensure representativeness and accuracy.

Diagnostics of Prior Betas: The following plots illustrates the convergence of chains, autocorrelation, shrink factor and density of the non-informative prior distribution:

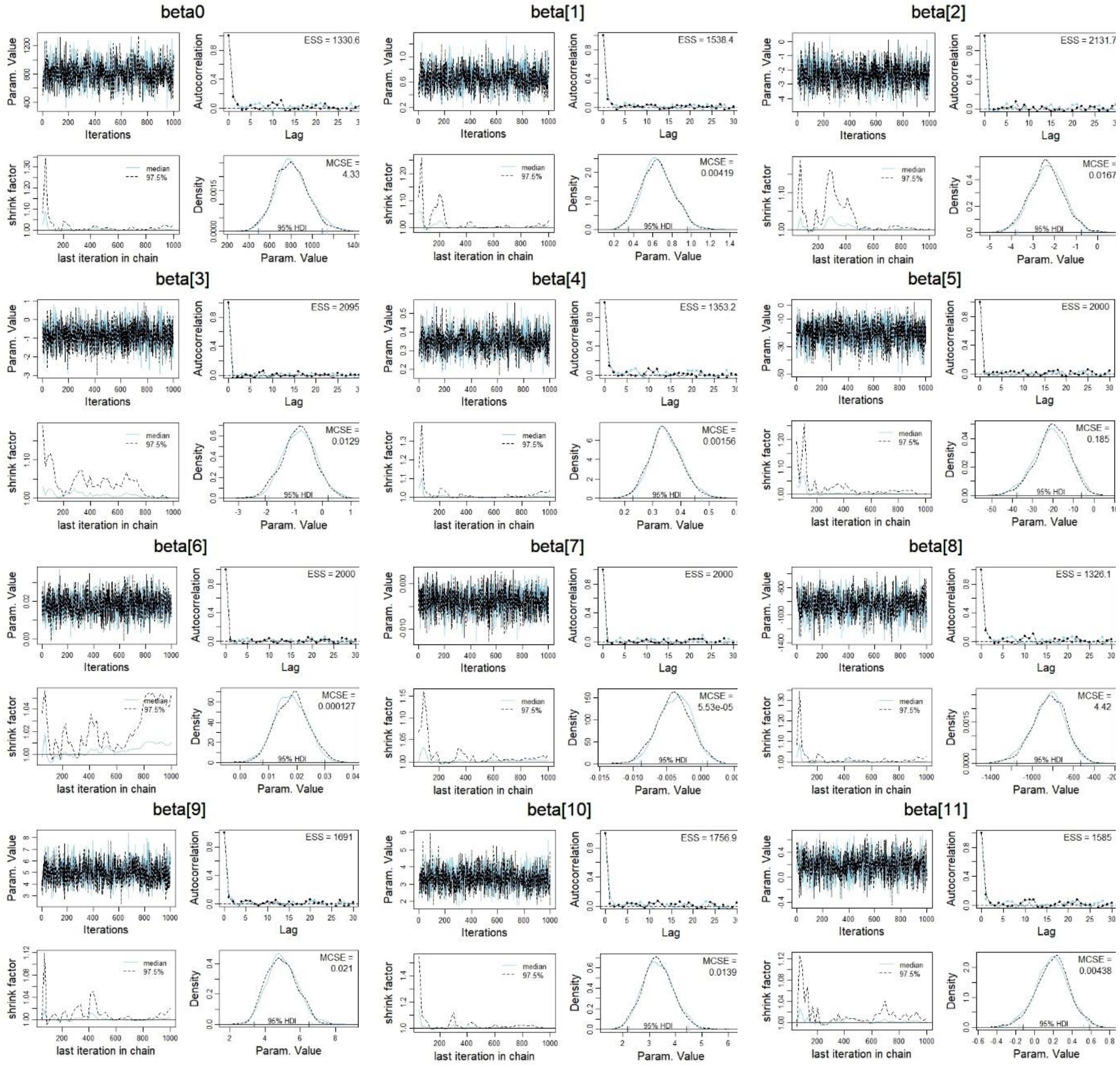


Figure 8: Diagnostic plots for prior betas

- **MCMC representativeness:**
  - **Chain convergence**: each pair of chains has a sufficient amount of convergence around the mean values of the posterior distribution. There is no sign of seasonality present. Thus, the larger burn-in steps work effectively in dropping the iterations at the start of the chain's generations.
  - **Density plot**: there are illustrations of the well overlapping of the two chains for each beta, around the parameter's values. Although they are not 100% overlapped with each other, they are considered acceptable given the efficiency of chain generation purpose.
  - **Shrink factor**: the shrink factors are well below the 1.2 value for all betas. After the further thinning steps method, beta 8's shrink factor had been reduced to an acceptable

level. Thus, it is not required to generate MCMC chains with larger settings as the further thinning steps method has proven to be both effective and efficient.

- **MCMC accuracy:**
  - **Autocorrelation:** the plots illustrate that no significant autocorrelations are presented that would impact the accuracy of the chains. Although some minor autocorrelations are found in some of the betas such as beta 1, beta 4, beta 8 etc., these minor autocorrelations are not particularly significant and are not to be concerned about. This is because each plot illustrates high effective sample size (ESS), which is upwards of several thousands.
  - **Monte Carlo Standard Error (MCSE)**: the MCSE of the diagnostic plots are very low or close to 0, which are indications of accurate chains generations for all betas.

- **MCMC efficiency:** given that this MCMC chain generation had an elapsed time of approximately 9 hours, which produced representative and accurate results, it could be inferred that the MCMC settings were indeed efficient. These settings with the use of further thinning steps were considerably more efficient than the larger MCMC settings with 5,000 saved-steps and absence of further thinning steps.

Diagnostics for Predictions: As there are 1,470 classification predictions to be made using the trained model, it is not efficient to conduct and discuss diagnostic testing for all classification predictions for the succinct purposes of this report. Thus, a handful of predictions will be randomly selected for diagnostic testing.

The following plots illustrates the convergence of chains, autocorrelation, shrink factor and density of the predictions produced by the non-informative prior distribution:

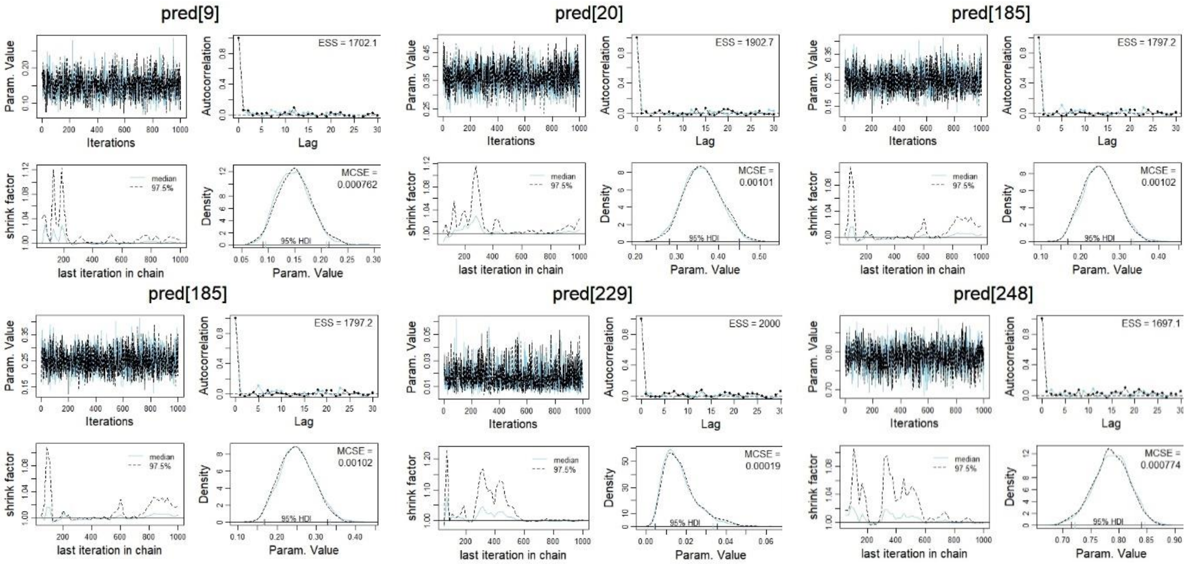


Figure 9: Sample of diagnostic plots for the predictions of non-informative prior

- **MCMC representativeness:**
    - **Chain convergence**: the chain convergence plot illustrates that there is no seasonality occurred in any sections of the iterations. The two chains are well converged around the mean values of the posterior distribution, indicating representativeness. Thus, it could be inferred that the large burn-in steps were effective in dropping steps prior to convergence, and the large adaptation steps worked well in training the model.
    - **Density plot**: The two chains of each density plot overlap with each other, with peaks around the parameter values. These density plots suggested that the prediction of classification of the white wines are representative, and thus, they are valid to be used for hypothesis tests.
    - **Shrink factor**: Predictions have shrink factor values of less than 1.2 which suggests that the chains are well representative, and valid to be used for classifications.

- **MCMC accuracy:**
    - **Autocorrelation:** the plots illustrate that no significant autocorrelations are presented that would impact the accuracy of the chains. Each plot illustrates high effective sample size (ESS), which is upwards of several thousands.
    - **Monte Carlo Standard Error (MCSE)**: the MCSE from the diagnostic tests are close to 0, which are indications of accurate chains generations.

4.2.2 Hypothesis testing and Bayesian estimates for coefficients

Given that the diagnostic tests illustrated that the generated MCMC chains are accurate and representative, they are valid to be utilised to obtain the posterior distributions. The following is the hypothesis statement used to test whether the independent variables are statistically significant to the classification model:

$$\mathbf{H_0}: \beta i = 0$$

$$\mathbf{H_A}: \beta i \neq 0;$$

Where *i* represents each independent variables within the model, *i = {1,2,3,4,5,6,7,8,9,10,11}*

The following is the MCMC plot that is used to determine the significance of the betas:

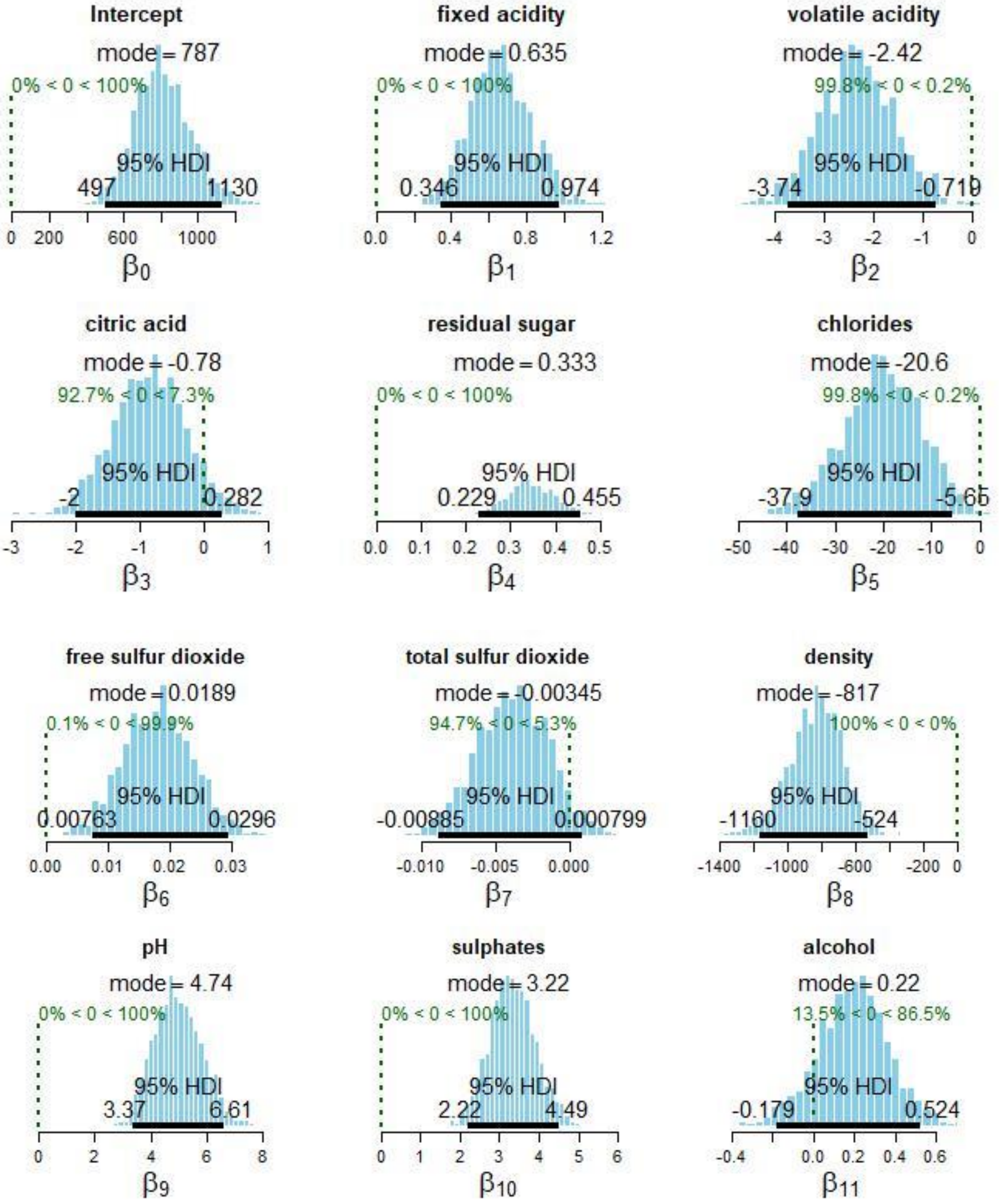


Figure 10: PlotMCMC for non-informative posterior distribution

The followings are discussions of Bayes' interval (HDI) that determine the significance of each of the independent variables:

- **Fixed acidity**: figure 8 illustrates that the fixed acidity has statistically significant contribution to the classification model. This is because the Bayes' interval of this independent variable does not include the value of 0. Thus, the probability that the beta of fixed acidity takes the values of 0.346 and 0.974 is 0.95.
- **Volatile acidity**: like the fixed acidity, the volatile acidity predictor variable is statistically significant as the value of 0 is not included and is far from the Bayes' interval. It is noted that this independent variable has an inverse relationship with the model, as its point estimate and Bayes' interval take negative values. The probability of volatile acidity variable assumes the values of -0.719 and -3.74 is 0.95.
- **Citric acid**: the probability that the beta for citric acid takes the values between -2 and 0.282 is 0.95. Although the Bayes' interval includes the values of 0, the probability that it takes a value of 0 is significantly low as 0 is located towards the end of the posterior distribution. Thus, it could be concluded that citric acid is significant towards the classification model.
- **Residual sugar**: the probability that the beta for residual sugars takes the values between 0.229 and 0.455 is 0.95. As 0 is not located within this Bayes' interval, it could be inferred that residual sugar is a significant independent variable towards the classification model. It is noted that this independent variable's posterior distribution as a small range, which is like the sample data.
- **Chlorides**: this predictor variable has an inverse relationship with the model, similarly to the volatile acidity. Chlorides has a significant distribution to the model as its Bayes' interval takes the values between -37.9 and -5.65, with a probability of 0.95. Unlike the residual sugar above, Chlorides has a considerably larger range.
- **Free sulfur dioxide**: the free sulfur dioxide independent variable has a Bayes' interval between 0.00763 and 0.0296 with a probability of 0.95, which does not include 0 within the interval. Thus, it could be inferred that this independent variable has statistically significant contribution to the classification model. Furthermore, this variable has longer tailed on both sides, which is a result of the outliers within the sample data.
- **Total sulfur dioxide**: this independent variable includes the value of 0 within the Bayes' interval. However, as 0 is located towards the upper tail of the 95% HDI, it could be deemed that it total sulfur dioxide has significant contribution to the model. The probability that its beta takes a value of -0.00885 and 0.000799 is 0.95.
- **Density**: the probability that the density beta has Bayes' interval between -524 and -1,160 is 0.95. This Bayes' interval is substantially far from the value of 0, thus, the density has statistically significant contribution to the classification model.
- **pH level**: pH level has statistically significant contribution to the classification model as its Bayes' interval does not include 0. Furthermore, pH level has a probability of 0.95 in taking a value between 3.37 and 6.61.

- **Sulphates**: the probability that Sulphates takes a value of 2.22 and 4.49 is 0.95. As its Bayes' interval does not include 0, it could be inferred that this independent variable has significant contribution to the classification model.
- **Alcohol**: it is reasonable to assume that alcohol level may have significant contribution to the quality of alcoholic drinks such as white wines. However, the Bayes' interval of this independent variable includes the value of 0, and the probability that its beta takes a value of -0.179 and 0.524 is 0.95. However, the probability that it takes the value of 0 is low as 0 is located towards the lower bound of the HDI. Thus, it could be concluded that alcohol level is a significant variable.

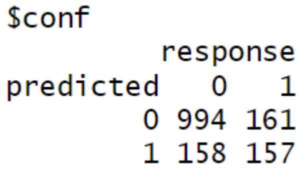
Bayesian point estimates: As discussed above, there is evidence to support that all the independent variables have significant contributions, thus, they are included in the model. The logistic regression model could be written as follows:

Y = logistics (787 + 0.635 Fixed acidity – 2.42 Volatile acidity – 0.78 citric acid + 0.333 residual sugar – 20.6 chlorides + 0.0189 free sulfur dioxide – 0.00345 total sulfur dioxide - 817 density + 4.74 pH + 3.22 sulphates + 0.22 alcohol)

It is concluded that the full logistics model could be obtained by the non-informative prior, and omission of insignificant independent variables is not required. It is noted that the intercept value is large, as it is mostly influenced by the density. Although the model suggests one unit change of an independent variable, independent variables such as 'density' would only change by a fraction of the unit as suggested by the descriptive looks.

4.2.3 Predictive check

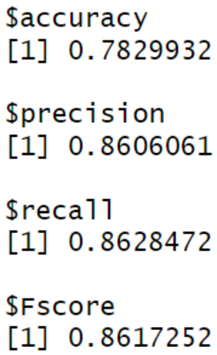
The Confusion matrix is a predictive check method that is used to determine the effectiveness of the model on generating accurate classification results. The following figure illustrates JAGS's confusion matrix for the non-informative prior:



```
$conf
         response
predicted   0   1
        0 994 161
        1 158 157
```

Figure 11: Confusion matrix of non-informative prior

The coefficient matrix illustrates that the non-informative prior model produced 994 *true negatives* and 157 *true positives.* In other words, the model had accurately classified 994 "Not high quality" and 157 "High quality" white wines respectively. This classification result is equivalent to approximately 78.3% accuracy.



```
$accuracy
[1] 0.7829932

$precision
[1] 0.8606061

$recall
[1] 0.8628472

$Fscore
[1] 0.8617252
```

Figure 12: Accuracy, precision, recall and F-score for non-informative prior

Classification results vs expectation: these classification results are aligned with the expectation discussed earlier in this report. Although the non-informative prior model was not able to accurately classify 100% of the white wines, a large portion (78.3%) of those were accurately classified. This indicates that the non-informative model is reasonably effective, given that no prior expert's information was incorporated into the model. As discussed, the remaining 30% of the sample data used for prediction may have some different characteristics than the 70% used to train the model. Hence, the model could not make complete accurate classifications for the wines.

Furthermore, Figure 12 above suggests that the non-informative model produced a precision score of 86.06%. This indicates that the model was above to identify and classify the majority of "high quality" white wines. The other 13.94% of the other "high quality" white wine had not been classified due to potential reasons such as outliers or being closely resemble to "not high quality" wines in characteristics.

4.3 Informative

4.3.1 Diagnostic tests for informative prior

The diagnostic checks for the informative prior model were conducted using the same MCMC settings as those applied in the non-informative case. However, it is worth noting that the MCMC diagnostics for the informative prior run indicate satisfactory convergence, and thus, no additional thinning steps are performed.

All the MCMC diagnostic plots for the non-standardized parameters show results similar to those of the standardized ones. Therefore, this study only presents the standardized results for this part. The total runtime was approximately 5 hours. The MCMC diagnostics based on the adjusted priors of standardized parameters and a set of predictions selected at random (Figure 13 and Figure 14) demonstrate the representativeness, accuracy, and efficiency of the MCMC chains.

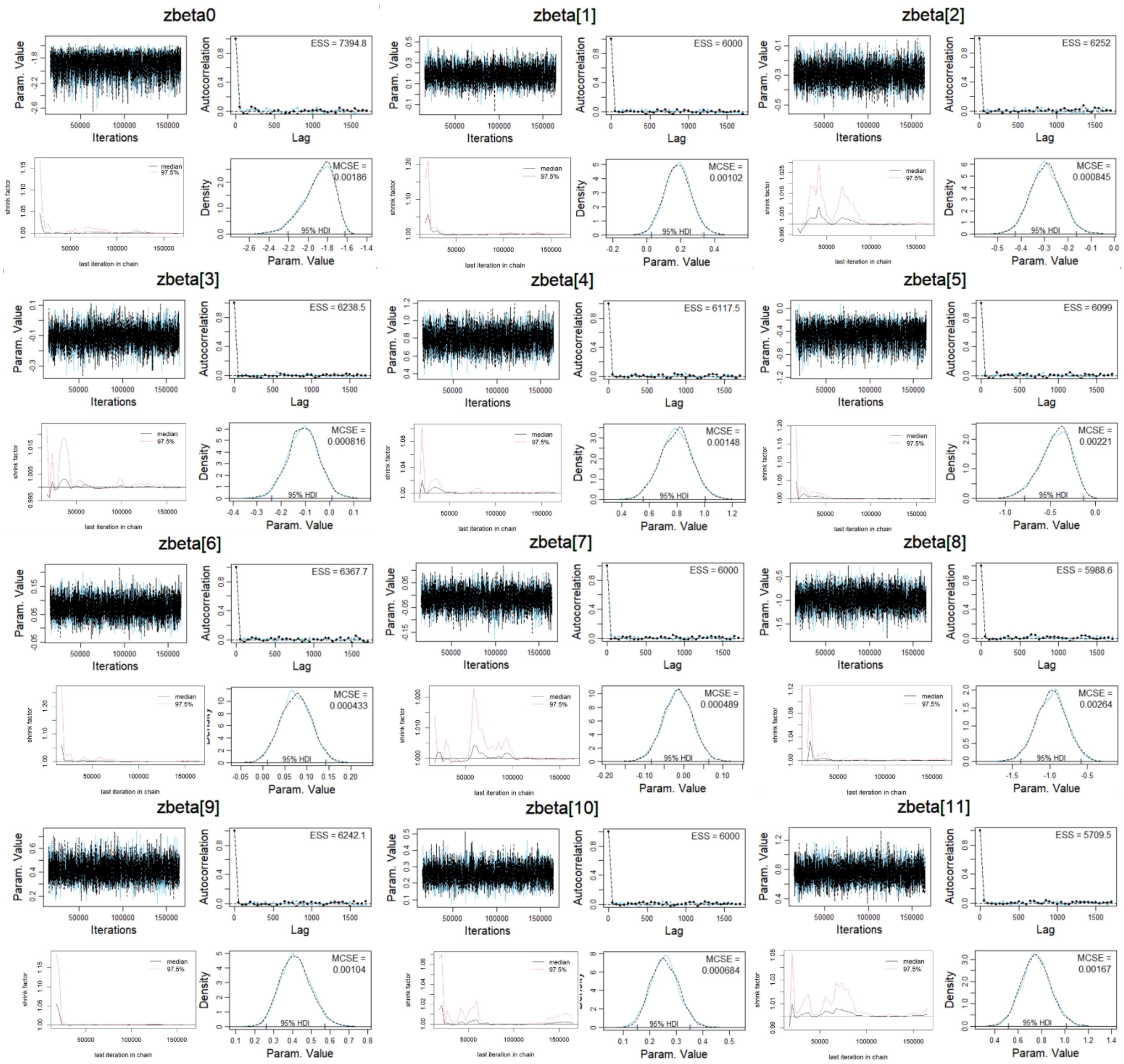


Figure 13: MCMC diagnostics for posterior estimates based on the informative priors

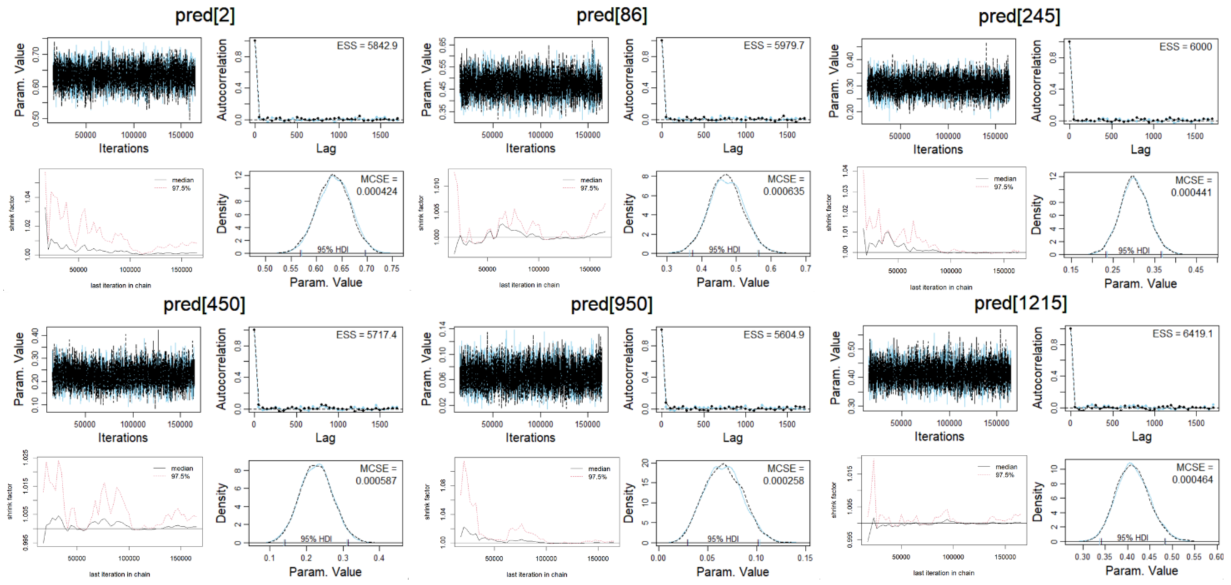


Figure 14: MCMC diagnostics for posterior estimates based on the informative priors of predictions

- MCMC Representativeness:

- **Chain convergence:** The chains show good convergence around the posterior means, with no seasonality observed. The use of large burn-in steps effectively discarded initial iterations, allowing the chains to stabilize.
- **Density plot:** The chains overlap well around the parameter values, which is sufficient for the purpose of chain generation.
- **Shrink factor:** The shrink factors for the median of all zbetas are below 1.1, with most of the 97.5% intervals also under 1.1. This confirms that the chains have converged to an acceptable level.

- MCMC Accuracy:

- **Autocorrelation:** No significant autocorrelation was found that would affect accuracy. Minor autocorrelations in some zbetas are negligible given the high effective sample size (ESS).
- **Monte Carlo Standard Error (MCSE):** The MCSE values are near zero, indicating highly accurate chains.

- **MCMC Efficiency:**

The MCMC run, which took around 5 hours, produced accurate and representative results, demonstrating that the settings used were efficient.

4.3.2 The Bayesian estimates and hypothesis testing for coefficients

The logistic regression model can be written as:

$$Y = logistic\ (415 + 0.297\ Fixed\ acidity - 2.86\ Volatile\ acidity - 0.863\ Citric\ acid + 0.201\ Residual\ sugar - 16.9\ Chlorides + 0.00575\ Free\ sulfur\ dioxide - 0.000112\ Total\ sulfur\ dioxide - 409\ Density + 2.95\ pH + 2.4\ Sulphates + 0.556\ Alcohol)$$

Figure 14 and Figure 15 present the posterior distributions of model parameters, providing deeper insights into how each physicochemical property affects the classification of wine quality. In addition, using the 95% highest density interval (HDI), hypothesis testing was conducted to assess the significance of each coefficient.

H_0: $\beta_i = 0$ (The independent variable does not significantly affect the success rate)

H_1: $\beta_i \neq 0$ (The independent variable significantly affects the success rate)

Where i represents the effect of a unit change in the corresponding independent variable, i = {1,2,3,4,5,6,7,8,9,10,11}

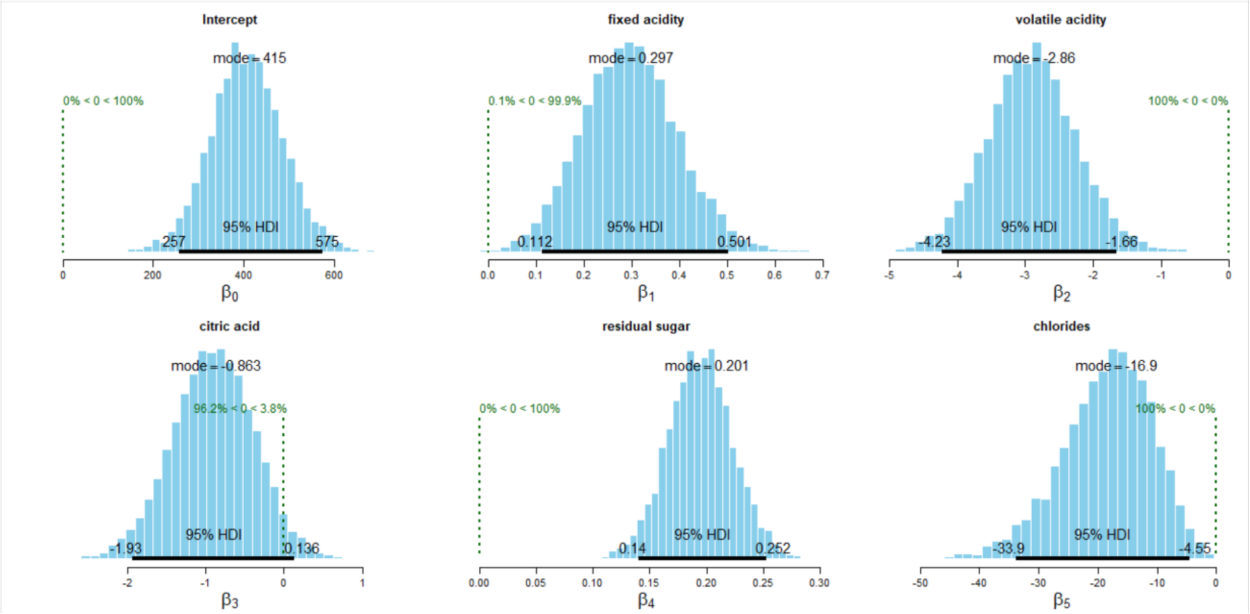


Figure 15: Posterior distributions of coefficients based on the informative priors (1)

The key details are as follows:

- Fixed Acidity (β_1): The posterior distribution of β_1 has a mode of 0.297 with a 95% HDI ranging from 0.112 to 0.501. This suggests that a one-unit increase in fixed acidity increases the odds of the wine being classified as high quality by approximately exp (0.297) ≈ 1.35 times. Additionally, since the HDI does not contain zero, the null hypothesis is rejected, meaning fixed acidity has a significant positive impact on wine quality.

- **Volatile Acidity (β_2):** The mode of -2.86, with a 95% probability that it lies between -4.23 and -1.66, indicates that a one-unit increase in volatile acidity reduces the odds of high-quality wine by approximately exp (-2.86) ≈ 0.057 times. Since zero is not within the 95% HDI, the null hypothesis is rejected, confirming that this effect is statistically significant.

- Citric Acid (β_3): Citric acid shows a negative relationship with the likelihood of high wine quality, indicated by a mode of -0.863 and a 95% HDI spanning from -1.93 to 0.136. A unit increase in citric acid decreases the odds of high-quality classification by about exp (-0.863) ≈ 0.42. Although the HDI slightly overlaps with zero, there's substantial evidence of a significant negative effect.

- **Residual Sugar (β_4)**: The posterior distribution of residual sugar has a mode of 0.201 with a narrow 95% HDI ranging from 0.14 to 0.252. This suggests that a one-unit increase in residual sugar increases the odds of the wine being classified as high quality by approximately exp (0.201) ≈ 1.22 times. Additionally, since the HDI does not contain zero, the null hypothesis is rejected, meaning fixed acidity has a significant positive impact on wine quality.

- **Chlorides (β_5)**: The mode of -16.9, with a 95% probability that it lies between -33.9 and -4.55, indicates that each unit increase in chlorides reduces the change of being high-quality classification by 4.58 x 10^{-8} times and the wide HDI reflects the variability in this effect. As the HDI excludes zero, the null hypothesis is rejected.

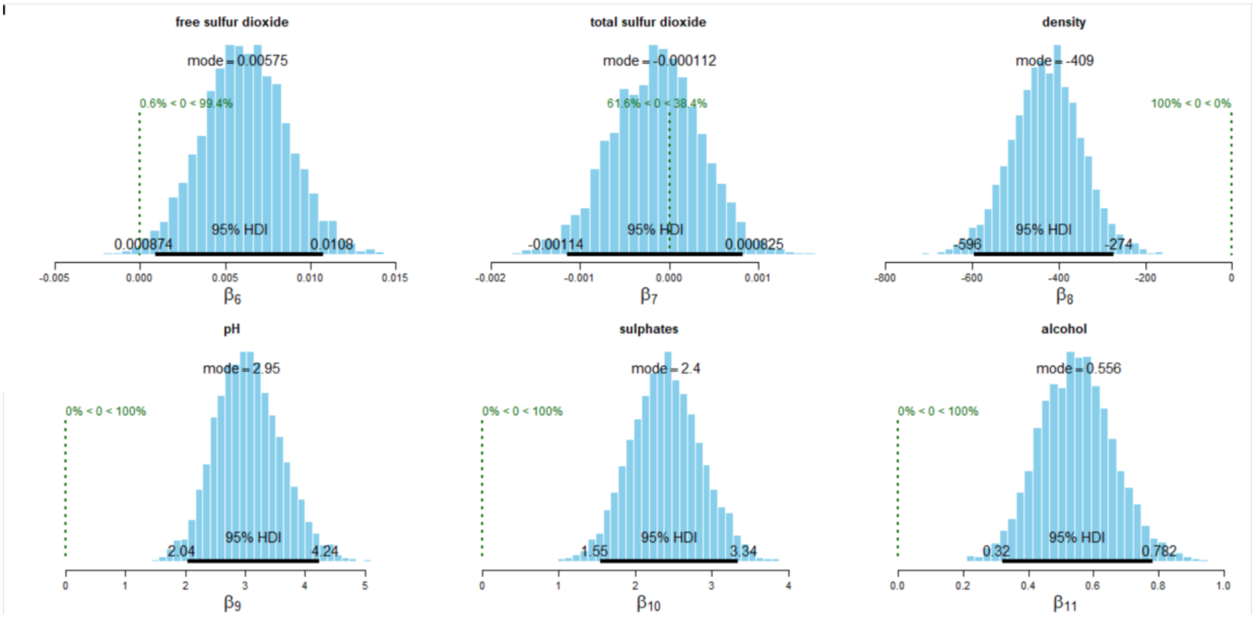


Figure 16: Posterior distributions of coefficients based on the informative priors (2)

- **Free Sulfur Dioxide (β_6)**: The mode of 0.00575 with a 95% HDI ranging from 0.000874 to 0.0108 indicates that free sulfur dioxide has a small yet significant positive effect on wine quality, raising the odds by exp (0.00575) ≈ 1. The HDI excludes zero, confirming the effect is significant.

- **Total Sulfur Dioxide (β_7)**: The mode of -0.000112 and a 95% HDI that includes zero indicate no strong evidence of total sulfur dioxide influencing the likelihood of high wine quality. As a result, the null hypothesis cannot be rejected.

- **Density (β_8)**: Density shows a strong negative effect, with a mode of -409 and an HDI from -596 to -274. A unit increase in density decreases the odds of high-quality classification by about exp (-409) ≈ 2.36 x 10^{-178}. Since zero is not part of the HDI, the null hypothesis is rejected.

- **pH (β_9)**: pH has a positive effect, with a mode of 2.95 and a 95% HDI ranging from 2.04 to 4.24. This suggests that a one-unit increase in pH increases the odds of the wine being classified as high quality by approximately exp (2.95) ≈ 19.11 times. Additionally, since the HDI does not contain zero, the null hypothesis is rejected, meaning pH has a significant positive impact on wine quality.

- **Sulphates (β_{10})**: The mode of 2.4, with a 95% probability that it lies between 1.55 and 3.34, indicates that a one-unit increase in sulphates raises the odds of high-quality classification by exp (2.4) ≈ 11.02 times. Since zero is not within the 95% HDI, the null hypothesis is rejected, confirming that this effect is statistically significant.

- **Alcohol (β_{11})**: Alcohol shows a positive relationship with the likelihood of high wine quality, indicated by a mode of 0.556 and a 95% HDI spanning from 0.32 to 0.782. Each unit increase in alcohol boosts the odds by roughly exp (0.556) ≈ 1.74, and the exclusion of zero from the HDI leads to the rejection of the null hypothesis.

The posterior analysis provides clear evidence that most features have a significant impact on wine quality, except for total sulfur dioxide, which shows no strong influence. Features such as pH, sulphates, alcohol, fixed acidity, residual sugar and free sulfur dioxide positively contribute to higher wine quality, while citric acid, volatile acidity, chlorides and density exert significant negative effects. pH and sulphates, in particular, stand out as key factors where the probability of a wine being classified as high-quality increases significantly as their levels rise. In contrast, citric acid emerges as one of the strongest negative predictors, where higher levels reduce the likelihood of high-quality classification. The feature importance results provide winemakers with valuable insights to optimize the production process, focusing on enhancing positive attributes like pH and sulphates, while reducing detrimental factors such as citric acid.

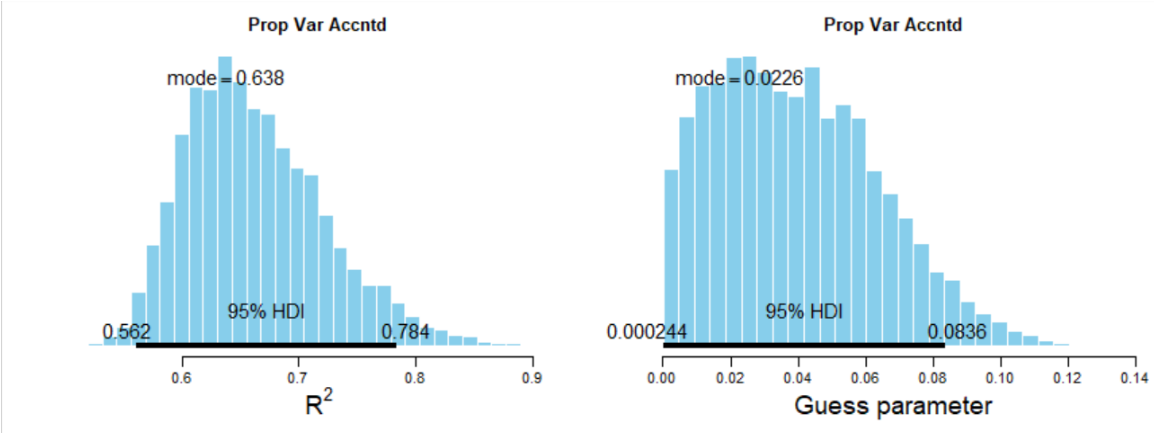
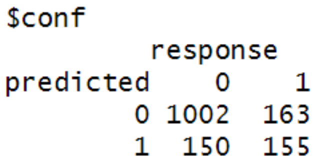


Figure 17: Posterior distributions of R^2 and guess parameter based on the informative priors

Moreover, the posterior distribution for R^2 has a mode of 0.638, with an HDI ranging from 0.562 to 0.784, indicating that the model explains a substantial portion of the variance in wine quality classification. Additionally, with median of 0.0226, the guess parameter distribution implies that nearly 98% of the times the success probability is specified by the model and 2% of the times it's impacted by the outliers and a random guess has been done.

4.3.3 Predictive check

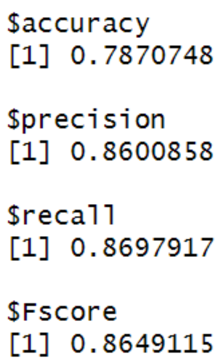
The confusion matrix reveals that the model accurately classifies a substantial number of both high and low-quality wines, with 1002 true negatives and 155 true positives. However, there are some misclassifications, including 163 false positives and 150 false negatives.



```
$conf
         response
predicted    0    1
        0 1002  163
        1  150  155
```

Figure 18: Confusion matrix showing predicted and actual wine quality classifications

This indicates that while the model is generally reliable, achieving an overall accuracy of approximately 78.7%, further improvements could be made by fine-tuning the decision threshold or incorporating additional features to enhance its classification performance.



```
$accuracy
[1] 0.7870748

$precision
[1] 0.8600858

$recall
[1] 0.8697917

$Fscore
[1] 0.8649115
```

Figure 19: Performance metrics for wine quality prediction mode

The model's predictive performance demonstrates a solid balance between precision and recall, as seen in the accuracy of 78.7%. While the precision of 86% suggests that most wines predicted as high quality are correctly classified, the 14% false positives indicate that a portion of lower-quality wines were incorrectly classified as high quality. The recall of 87% is strong, meaning the model successfully identifies a majority of actual high-quality wines, but the 13% of high-quality wines that were missed as false negatives show room for improvement in identifying truly exceptional wines. The F1 score of 86.5% illustrates an overall balanced model, combining both the ability to minimize false positives and effectively capture true positives.

5. Sensitivity Analysis

5.1 Non-informative

A sensitivity analysis was performed on the non-informative prior distribution to determine the validity of the model discussed above. For consistency purposes, the MCMC settings run was kept constant at 5,000 burn-ins, 10,000 adaptation steps, 2 chains, 50 thinning steps, and 3,000 saved steps. Furthermore, the 3 additional thinning steps were also applied to reduce the autocorrelation, similarly to the initial non-informative run.

Although the settings and JAGS model were kept primarily the same, larger variances were applied around the mean of the normal prior distributions. The purpose of applying larger variances is to determine whether the initial specifications for the prior, Normal ~ (0, 2), were large enough to have no influences on the posterior distributions. Thus, a specification of Normal ~ (0, 12) was applied for the sensitivity analysis.

The validity of the initial specifications is determined by comparing the two posterior distribution results. It suggests that the initial specifications is valid (adequately large variances) if they yielded similar results. However, it is noted they would not yield the exact results, as MCMC runs would produce different results due to randomness of chain generations. The following plots are the posterior distributions from the sensitivity analysis run:

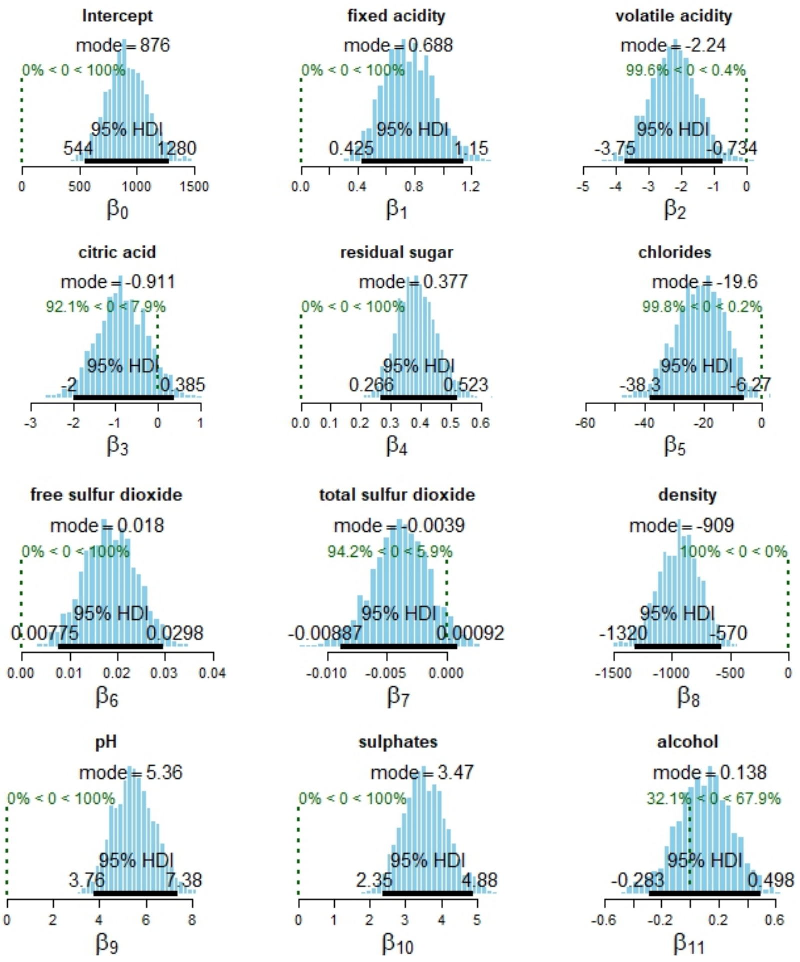


Figure 20: Sensitivity analysis for non-informative prior distribution

The diagnostic tests for the sensitivity analysis had been performed and concluded that all MCMC chains are representative and accurate. However, they are not discussed for succinct purposes of this report. Figure_ shows that sensitivity analyses' posterior distributions yielded similar results to initial non-informative runs, despite having larger prior variances around the means. Although there are slight differences, they are negligible as they are results of randomly from the MCMC chain generations.

This concludes that the specifications of the initial non-informative is valid. In other words, the prior distributions N ~ (0, 2) had large enough variances that allowed the likelihood to dominate the posterior distributions.

5.2 Informative

This study conducts a sensitivity analysis to identify the prior variance values that lead to either informative or non-informative prior distributions. Gibbs sampling was carried out consistently across different variance settings (Var = 0.1, 0.5, 1, 1.5, 2, 3, 5), using the specified prior information to observe how each setting influences the posterior estimates.

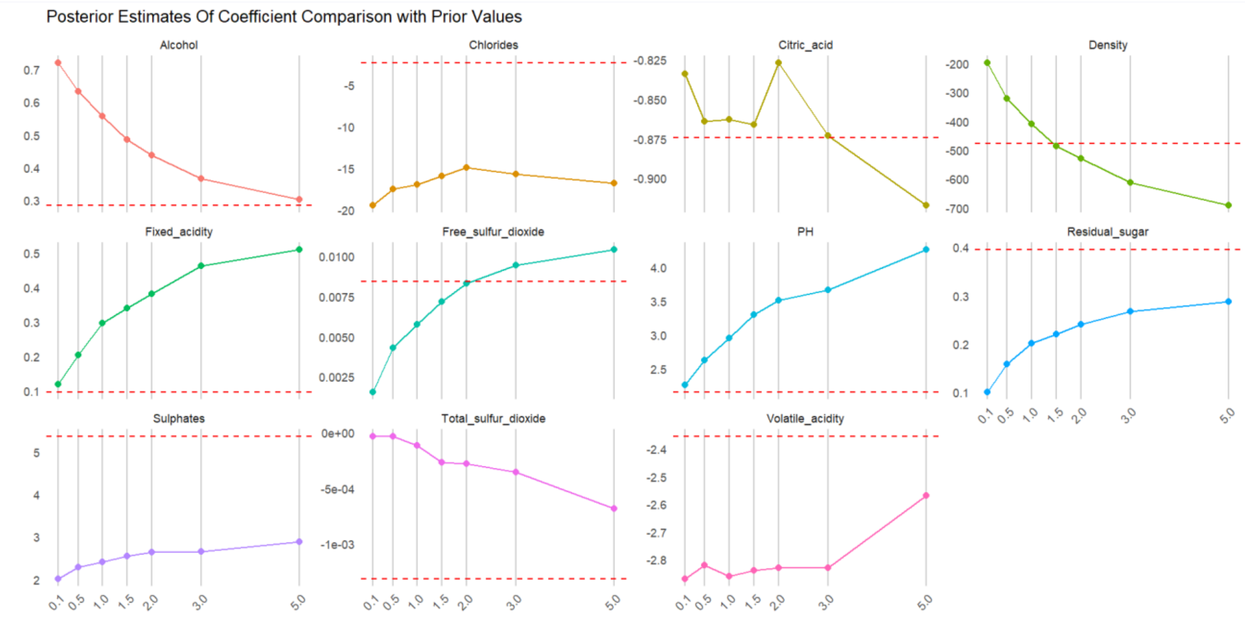


Figure 21: Posterior sensitivities of coefficient estimates for various parameter values. The x-axis showing variance values, the y-axis showing posterior estimates, red dashed lines indicating prior values

Figure 21 shows a sensitivity analysis of posterior coefficient estimates as we vary prior variances. The x-axis represents different parameter values, while the y-axis shows the posterior estimates, with each red dashed line marking the prior for a specific coefficient. From the figure, we can see that when the prior variance is set to 1.5, the posterior estimate for Density is close to the prior, suggesting a strong alignment between prior beliefs and observed data. Similarly, Free sulfur dioxide aligns closely with the prior at a variance of 2, and Citric acid shows similar alignment at a variance of 3. Thus, choosing a prior variance of 1 provides a balanced approach where prior beliefs have a meaningful influence on the posterior estimates without overwhelming the new insights provided by the data.

IV. Conclusion

The Bayesian robust logistic regression model provides valuable insights into the classification of white wine quality, allowing for a nuanced understanding of the influence of various physicochemical properties. Significant features such as pH, sulphates, alcohol, fixed acidity and residual sugar were identified as positive indicators of higher quality, while volatile acidity, citric acid, chlorides and density showed negative impacts.

The predictive check results indicate that both the non-informative and informative models demonstrate strong predictive accuracy and high precision, confirming the effectiveness of the logistic regression framework for classifying wine quality, with accuracy rates of 78.3% and 78.7%, respectively. The informative model, enhanced by expert knowledge on predictor importance, achieved slightly better accuracy, highlighting the advantages of emphasizing key features. While both models provide reliable classifications, the informative model's use of domain insights offers a slightly more refined performance, making it especially valuable for industry-focused applications. Meanwhile, the non-informative model serves as a robust, unbiased approach, providing a solid foundation for classification without prior assumptions.

Furthermore, the sensitivity analysis indicates that the model is robust to variations in prior assumptions, providing confidence in the reliability of the findings.

The findings from this research not only help predict wine quality with high accuracy but also highlight specific chemical properties that producers can adjust to enhance quality. By incorporating both non-informative and informative priors, the model offers a comprehensive view of the factors influencing quality, positioning it as a valuable tool for the wine industry's move toward data-driven quality assurance.

References

Cortez, P. et al (2009). Modeling wine preferences by data mining from physicochemical properties. Semantic Scholar. https://www.semanticscholar.org/paper/Modeling-wine-preferences-by-data-mining-from-Cortez-Cerdeira/bf15a0ccc14ac1deb5cea570c870389c16be019c

Cortez, P. et al (2009). *Wine Quality [Dataset]*. Machine Learning Repository. https://archive.ics.uci.edu/dataset/186/wine+quality

Demirhan, H., Demirhan, K. A Bayesian approach for the estimation of probability distributions under finite sample space. *Stat Papers* **57**, 589–603 (2016). https://doi.org/10.1007/s00362-015-0669-z

Demirhan, H. (n.d.) Beta Distribution Specified by Mean and Concentration. [Shinyapps.io]. https://rmitsam.shinyapps.io/beta_3/

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

Kruschke, J. K. (2015). Doing Bayesian Data Analysis, Second Edition: A Tutorial with R, JAGS, and Stan. Academic Press / Elsevier.