**Classifying the Severity of Vent-flow Events in Alberta’s Oil and Gas Wells**

**DATA 606 – PROJECT REPORT**

Jincy Thomas (30271046)

Joshua Quartey (30262911)

Megha Radhakrishnan Sanitha (30272516)

​

**Contents**

[**INTRODUCTION** 2](#_Toc200917969)

[**DATASET** 2](#_Toc200917970)

[**GUIDING QUESTIONS** 3](#_Toc200917971)

[**DATA PREPROCESSING** 3](#_Toc200917972)

[**EXPLORATORY DATA ANALYSIS** 4](#_Toc200917973)

[**ANALYSIS** 7](#_Toc200917974)

[**LOGISTIC REGRESSION - Assumptions** 7](#_Toc200917975)

[**LINEAR DISCRIMINANT ANALYSIS (LDA) - Assumptions** 11](#_Toc200917976)

[**Logistic Regression and LDA – Model** 14](#_Toc200917977)

[**QUADRATIC DISCRIMINATIVE ANALYSIS (QDA)** 19](#_Toc200917978)

[**QDA Model – (Flow Substance Included)** 32](#_Toc200917979)

[**CLASSIFICATION TREE** 39](#_Toc200917980)

[**CONCLUSION** 47](#_Toc200917981)

[**Future Improvement** 48](#_Toc200917982)

[**REFERENCES** 49](#_Toc200917983)

# **INTRODUCTION**

Surface casing vent flow (SCVF) and gas migration events are critical integrity issues in the oil and gas industry which can lead to unintended releases of methane and other hydrocarbons into the environment. Subsurface pressure forces up to the surface natural gas or liquids through the well casing causing environmental impacts such as greenhouse gas emissions, groundwater contamination, and surface blowouts (BC Energy Regulator n.d.).

The Alberta Energy Regulator (AER) maintains a daily‐updated Vent Flow and Gas Migration Report, recording well characteristics, gas composition, and measured gas volumes to help track and monitor these incidents (Alberta Energy Regulator 2024). Timely identification of wells prone to Surface Casing Vent Flow is essential due to the environmental and safety hazards risks they pose.

Although measures like inspection routines are in place, many vent‐flow incidents remain undetected until surface releases occur. Predictive analytics could complement current safety measures by improving their precision and effectiveness (Schiffner, Kecinski & Mohapatra 2021).

In this project, our goal is to use AER’s Vent Flow and Gas Migration dataset to

(a) Identify key well features that could be used to predict the severity of vent‐flow events

(b) Build and compare logistic, LDA, QDA and decision‐tree classifiers in predicting event severity.

(c) Deliver actionable insights based on our results, like decision rules or features that identify high‐risk wells, for field engineers and regulators to help reduce the frequency and impact of uncontrolled gas releases.

# **DATASET**

Data source: [AER General Well Data – Vent Flow and Gas Migration Report](https://www.aer.ca/data-and-performance-reports/activity-and-data/general-well-data)

Vent Flow and Gas Migration Report, a publicly available dataset for non-commercial, educational use provided by the Alberta Energy Regulator (AER) website, was used for our project. The license terms can be found [here](https://www.aer.ca/copyright-and-disclaimer). This dataset provides a complete record of oil, gas, oil sands and water wells in Alberta, Canada, that have reported surface casing vent flow (SCVF) or gas migration (GM) events. The data is provided in CSV format and has 39321 rows.

The columns which were used in our statistical analysis are:

1. **Classification**

Type: Categorical

Values: Serious, Non-Serious

Description: Regulatory classification of the event

1. **flow\_substance**

Type: Categorical

Values: Condensate, Crude Bitumen, Crude Oil, Gas, Hydrocarbon Liquid, Non-Saline, Not Converted, Other, Saline, Unusable Water, Usable Water, Waste

Description: Substance detected flowing

1. **flow\_rate\_m3\_day**

Type: Numerical

Description: Measured flow rate in cubic meters per day

1. **groundwater\_base\_mkb**

Type: Numerical

1. Description: Depth to groundwater base below Kelly Bushing in mKB **stabilized\_shut\_in\_pressure\_kPa**

Type: Numerical  
 Description: Pressure measured when the well is shut in

1. **source\_depth\_mkb**

Type: Numerical  
 Description: Depth of the source below Kelly Bushing in mKB

# **GUIDING QUESTIONS**

1. Which characteristics are best predictors for regulatory classification (Serious, Non-Serious)?
2. Can Linear Discriminant Analysis (LDA) or Quadratic Discriminant Analysis (QDA) effectively classify events based on the available features?
3. How accurately doour classification models predict the severity of events when evaluated with cross-validation and confusion matrix?

# **DATA PREPROCESSING**

The original dataset contained 39,319 rows and 16 columns, and the initial steps included identifying and handling missing values, trimming white spaces.

##### Handling Missing Values

Significant missingness was observed, particularly in key numerical columns:

* Flow Rate (m3/day) – 16,740 missing
* Stabilized Shut in Pressure (kPa) – 3,567 missing
* Source Depth (mkb) – 23,019 missing

After removing rows with any missing values, 9,750 clean rows remained for further analysis.

##### Dropping Irrelevant Columns

To focus on relevant operational and physical parameters, the following columns were removed:

* Report metadata (e.g., licence number, surface/bottom location, report/resolution dates).
* Licensee details and current licence status.

##### Column Name Standardization

Column names were cleaned by replacing spaces with underscores, and special characters were removed.

Example: Flow Rate (m3/day) → Flow\_Rate\_m3\_day

##### Categorical Feature Encoding

* Columns were converted to factor types.
* Classification (converted to two classes: Serious, Non-Serious)

##### Removal of Zero Rows

Rows where all numeric columns were zero were excluded to ensure meaningful analysis.

# **EXPLORATORY DATA ANALYSIS**

#### Data Summary

* Most cases are classified as Non-Serious (6,010), with 2,074 marked as Serious.
* Most incidents are of the Vent Flow type (7,847).
* The predominant flow substance is Gas.

#### Univariate Analysis

A graph of different types of data

AI-generated content may be incorrect.

Figure 1:Boxplots - distribution of each numerical variable

The boxplots show the distribution of four key numeric variables:  
**Flow Rate**, **Shut-In Pressure**, **Source Depth**, and **Groundwater Base Depth**.

All four variables are **right-skewed** and contain **extreme values (outliers)**, particularly in **Flow\_Rate\_m3\_day** and **Stabilized\_Shut\_In\_Pressure\_kPa**.

These outliers were **not removed or transformed**, as they represent **genuine observations** in the dataset and could be **critical for understanding serious incidents** or rare but important events.

Preserving these values helps maintain the **integrity and representativeness** of the data, especially in safety-critical domains where unusual readings might signal significant risks.

A collage of graphs

AI-generated content may be incorrect.

Figure 2: Histograms for each numerical variable

1. Flow\_Rate\_m3\_day

* Most flow rate values are clustered very close to 0, indicating most incidents have very low flow rates.

2. Stabilized\_Shut\_In\_Pressure\_kPa

* Similarly, shut-in pressure values are heavily concentrated at lower levels, with relatively few cases showing high pressure.

3. Source\_Depth\_mkb

* Most sources are located at very shallow depths, with a rapid drop-off in frequency as depth increases.

4. Ground\_Water\_Base\_mkb

* Groundwater base depths are also heavily right skewed, with most values under 2 meters below Kelly bushing.

#### Analysis of Categorical Variables

A graph of a number of data

AI-generated content may be incorrect.

Figure 3: Source Depth by Classification

The distribution is heavily right skewed, indicating that most incidents occur at shallow depths. Serious cases, while also concentrated at shallow depths, are more dispersed and appear more frequently at slightly deeper levels compared to non-serious cases.

A graph of a type distribution

AI-generated content may be incorrect.

Figure 4: Type by Classification

The data show that Vent Flow is the predominant incident type, accounting for the largest share of both Non-Serious and Serious cases. Gas Migration and Vent Flow/Gas Migration events are relatively infrequent, with the latter more likely to be classified as Serious. This suggests that while most incidents are not serious, particular attention should be paid to Vent Flow/Gas Migration events due to their higher proportion of seriousness.

A graph of different substances

AI-generated content may be incorrect.

Figure 5: Flow Substance by Classification

The gas is the primary flow substance involved in both Non-Serious and Serious incidents, highlighting its prevalence and risk potential.

# **ANALYSIS**

### **LOGISTIC REGRESSION - Assumptions**

Logistic regression assumes that:

* The outcome is a binary or dichotomous variable
* There is a linear relationship between the logit of the outcome and each predictor variables.
* There are no influential values (extreme values or outliers) in the continuous predictors
* There is no high intercorrelations (i.e. multicollinearity) among the predictors.

Making sure these assumptions are met by the data will help improve the accuracy of the model. In the following sections, we will describe how we diagnosed which of the assumptions were met and any potential problems in the data.

* Linearity assumption

Here, we check the linear relationship between the continuous predictor variables and the logit of the outcome. This can be done by visually inspecting the scatter plot between each predictor and the logit values.

Overall, none of the four numeric predictors shows a completely straight outcome logit-predictors trend, each plot has at least one bend or plateau. If the curve bends, the coefficient estimate for that predictor can be biased and the Wald tests unreliable

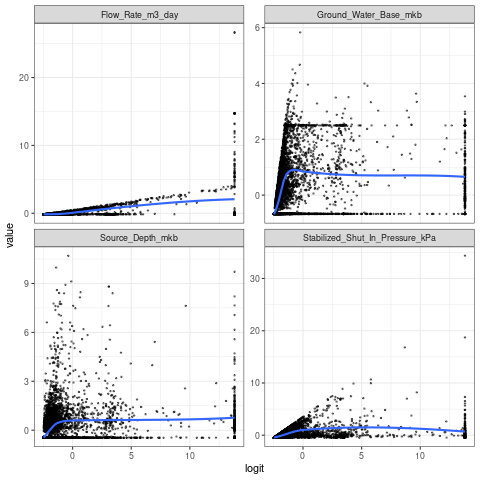


Figure 6: Basic logistic model linearity of log-odds plot

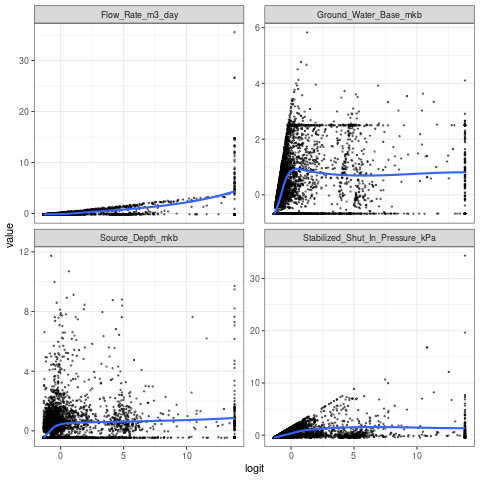


Figure 7: Weighted logistic model linearity of log odds plot

After including polynomial terms, the smoothed scatter plots show that variables Flow rate, Source depth, Stabilized shut in pressure, and pressure and ground water base are all quite linearly associated with the diabetes outcome in logit scale.

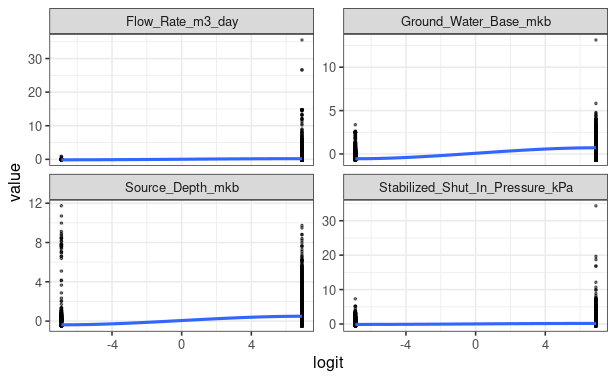


Figure 8: Polynomial logistic model linearity of log-odds plot

* Influential values

Influential values are extreme individual data points that can alter the quality of the logistic regression model.

##### Polynomial model

The most extreme values in the data can be examined by visualizing the Cook’s distance values. Observation #3521 has a Cook’s Distance well above every conventional cutoff (4 / n or 1). The two other cases (#7655, #9621) are orders of magnitude smaller compared with #3521.

From the residuals plot, one point has a leverage $\approx$ 0.78 and a large negative Pearson residual (far below $–1 × 10\_8$). This single observation is both high leverage and has an extreme residual and so could be highly influential on the fitted coefficients.

##### Base model

The plots show low-leverage points with large residuals but are not influential.

|  |  |
| --- | --- |
| Figure 9: Residual-leverage plot basic logistic model | Figure 10: Cook’s distance plot basic logistic model |
| Figure 11: Residual-leverage plot polynomial logistic model | Figure 12: Cook’s distance plot polynomial logistic model |

* Multicollinearity

Multicolinearity occurs when the data contains highly correlated predictor variables.

Multicollinearity is an important issue in regression analysis and is typically fixed by removing the concerned variables. It can be assessed using the variance inflation factors:

|  |  |
| --- | --- |
| Figure 13: Weighted logistic model VIF scores | Figure 14: Basic logistic model VIF Scores |

A VIF value exceeding 4 or 10 indicates a problematic amount of collinearity. In both models (base and polynomial), there is no collinearity, since all variables have VIFs below 4.

##### Outcome

At the end of the Logistic Regression assumptions diagnostics:

* Using a basic model trained on all predictors in the dataset results in a model that does not meet the linearity of predictors and logits of outcome assumption but meets the no outlier assumption and the multicollinearity assumption.
* Transforming the predictors by including polynomial terms produces a model that meets the linearity assumption, influential outliers and the multicollinearity assumption.

Given its clearer interpretation, cleaner diagnostics and reduced implementation risk, the weighted and basic logistic regression is the preferred choice over the polynomial alternative.

### **LINEAR DISCRIMINANT ANALYSIS (LDA) - Assumptions**

Linear Discriminant Analysis assumes the following:

* Predictors have the same variance-covariance matrix for each class. The variables have the same variability and correlation structure across the different groups.
* Each class follows a multivariate normal distribution.
* Observations are independent of one another.
* No multicollinearity among predictors

In the following sections, we check each of these assumptions on our data

* Multivariate Normality
* Multicollinearity
* Multivariate Homogeneity of Dispersions

##### Multivariate Normality

The multivariate Shapiro-Wilk test gave us a very small p-value, which suggests that we reject the null hypothesis, the predictors are not normally distributed.

H0: The data comes from a multivariate normal distribution

Ha: The data does not come from a multivariate normal distribution.

The p-value of the test is 0.0002692. Since this is a lot less than .05, we reject the null hypothesis of the test. We have enough evidence to say that the four variables in our dataset do not follow a multivariate normal distribution.

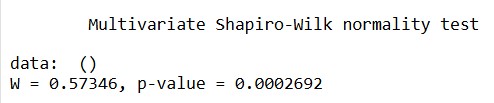


Figure 15: Multivariate Shapiro-Wilk test results

The energy test for multivariate normality also confirms our previous results. Because the test's p-value(< 2.2e-16) is < 0.05, we reject the null hypothesis and conclude that the predictors are not drawn from a multivariate normal distribution. This violates the multivariate normality assumption.

H0 : The data is drawn from a multivariate normal distribution

Ha : The data is not drawn from a multivariate normal distribution

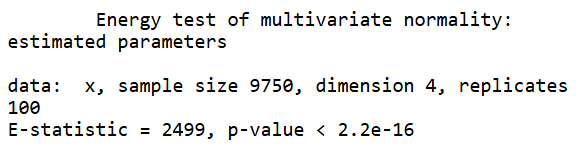


Figure 16: Multivariate normality Energy test results

##### Multicollinearity

Since the VIF scores for the numerical columns are all less than 4, we can conclude that there is no problem of multicollinearity between the predictors.



Figure 17: Multicollinearity test LDA

##### Multivariate Homogeneity/Homogeneity of Variance

One of the key assumptions of LDA is that the predictor variables have the same variance-covariance matrix for each class. The variables have the same variability and correlation structure across the different groups.

The code below checks whether wells grouped by Classification have similar multivariate variance (dispersion) across the selected well parameters. If the ANOVA is non-significant, the dispersion (i.e. multivariate variance) is considered homogeneous.

Since the data is not normally distributed, we use the Levene's test to test for multivariate homogeneity

H0: There is a significant difference between tested sample variances

Ha: There is no significant difference between tested sample variances

The null hypothesis was strongly rejected for every numeric predictor; the spread of each numeric variable differs between “Serious” and “Non Serious” wells. This violates the equal variances between classes assumption.

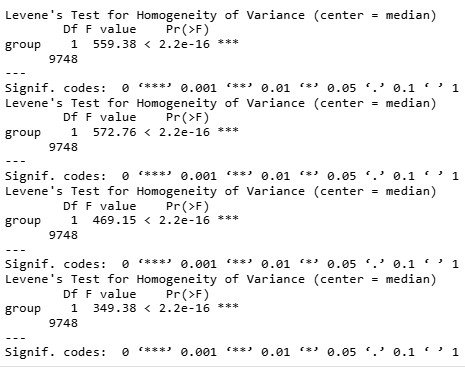


Figure 18: Levene’s test results

Comparing the ratios of the class variances, we see there is strong heteroscedasticity where ratios are above 3 to 4 times (Flow-Rate, Pressure). The “Serious” wells are far more variable, especially in flow rate (approximately 50 times) and shut-in pressure (approximately 8 times)

LDA assumes equal class-covariance matrices and so going forward with this technique would mean the model is more likely to mis-classify when one class is much more dispersed*.*

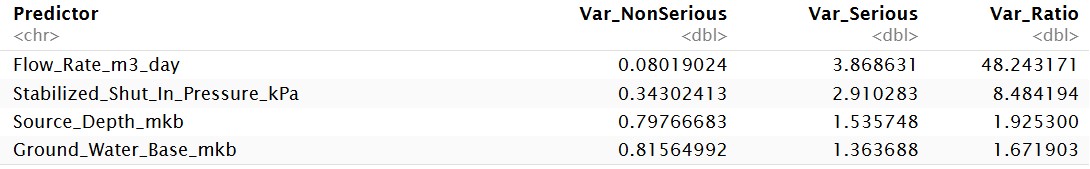


Figure 19: Class variance ratios

##### Outcome

The only assumption met for Linear Discriminant Analysis was the multicollinearity assumption. The results show that Linear Discriminant Analysis is not reliable here. Using this method would result in poorer accuracy, distorted posterior probabilities, and potential unfairness to the minority class.

The results from the assumptions check indicate that LDA’s linear boundary would be biased toward the more variable class("Serious"), and posterior probabilities and mis-classification rates can be badly skewed.

We will keep LDA in our analysis, mainly to serve as a basis for comparison with the expectation that it will under-perform.

### **Logistic Regression and LDA – Model**

*Model* building, training and evaluation

The models were each tuned and validated on the training set (75% stratified split) using stratified k-fold (k= 10) cross validation and then finally their performance on the test set was compared.

This training methodology was chosen because:

* By holding out a separate test set and not using it in any part of model training or hyperparameter tuning, this approach avoids data leakage, preserving the integrity of final model evaluation.
* The test set provides an unbiased estimate of the model’s generalization performance, as it is never seen during training or validation.
* Applying k-fold CV on the training set makes efficient use of data for tuning and performance assessment, especially when the dataset is small or moderate in size.

#### **Model Results and Interpretations**

##### **Logistic Regression**

###### Basic Logistic Regression

The baseline logistic model retained three of the four continuous predictors ("Flow Rate (m³ day⁻¹"), "Stabilized Shut-In Pressure (kPa)", and "Ground-Water-Base Depth (m kb)". Each predictor was highly significant (p-value < 0.001). The largest effect was "Flow Rate" (beta ≈ 3.21), indicating that a unit increase multiplies the odds of a "Serious" vent flow event by about 25 (e3.21).

Cross-validated performance showed a mis-classification rate of 17 % and AUC ≈ 0.77. The test set had almost identical (16%, AUC 0.80), confirming good generalisation. Specificity is very high (0.98) but sensitivity is modest (0.34), meaning the classifier correctly rejects most Non-Serious wells but misses two-thirds of the minority class. This imbalance-driven bias reflects the original class proportions (77 % Non-Serious), which pull the decision boundary toward the majority class and minimise false alarms at the expense of recall.

###### Weighted Logistic Regression

Applying inverse-frequency weights (w ≈ 1/0.77 for Non-Serious, 1/0.23 for Serious) shifts the intercept (+0.12) and slightly raises all slopes (e.g beta\_{flow} ≈ 3.52$) compared to the baseline model. The cross-validated sensitivity more than doubles to 0.61 while specificity is lower(0.82); balanced accuracy is higher(from 0.65 to 0.71). On the hold-out test set, sensitivity improves to 0.64, while specificity remains at 0.83, and AUC also remains at 0.80. The overall accuracy drop to 79% (mis-class ≈ 21%), but the model now identifies almost two-thirds of Serious wells. These results seem to indicate that weights are able to counteract class-imbalance bias in our case.

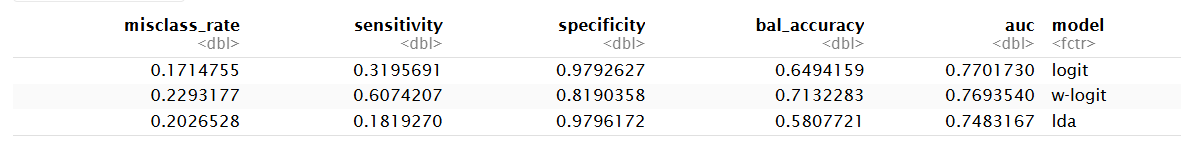


Figure 20: Cross-validation metrics

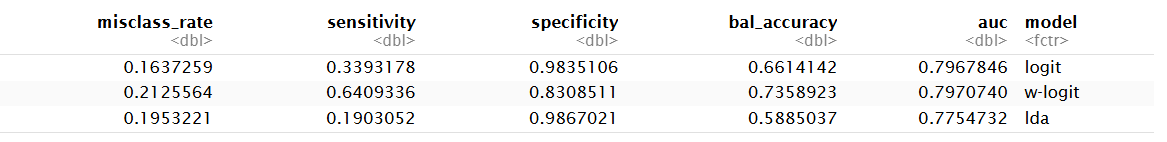


Figure 21: Test set metrics

##### **Linear Discriminant Analysis**

With only two classes, LDA produces a single discriminant axis The formula of the model is:

LD1 = 0.618∗Flow\_Rate\_m3\_day+ 0.587∗Stabilized\_Shut\_In\_Pressure\_kPa + 0.537∗Ground\_Water\_Base\_mkb

The plots show substantial overlap between classes on LD1. LDA’s linear boundary in each plot is driven mostly by "Flow Rate" and "Shut-In Pressure". "Ground-Water Base" rarely determines a prediction. The purple area containing many red “N” points illustrates why specificity is high (Non-Serious wells often kept) but sensitivity is low (Serious wells share the same region)

The strong dependence on Flow Rate is consistent across both ordinal and partition views, confirming it as the dominant LDA driver.

Because the decision surface is linear and based on population means, it cannot extract the more complex curved relationship visible in the logistic plots, which accounts for LDA’s inferior recall. The cross-validated balanced accuracy is 0.58 and test-set 0.59. There seems to be very low sensitivity (about 0.19) despite specificity of 0.98. The poor recall is likely due to the two violated assumptions:

(i) within-class normality, both densities are heavily right-skewed

(ii) equal covariance matrices, variance ratios across predictors favour the majority class.

As a result, LDA under-predicts "Serious" wells and offers no advantage over logistic alternatives.

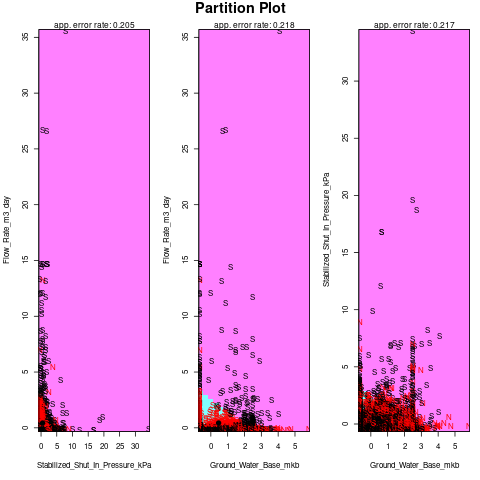


Figure 22: Final LDA model partition plot

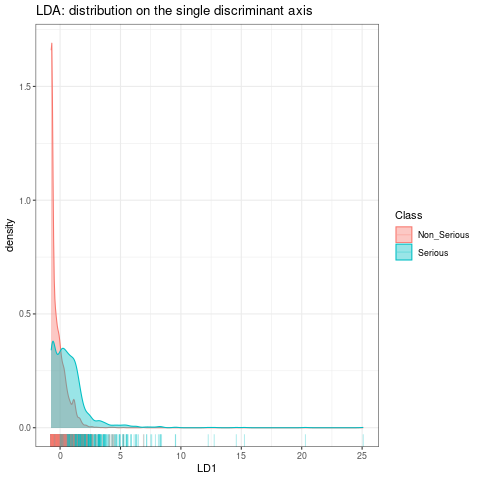


Figure 23: LDA ordinal plot

##### **Discussion of results**

###### **Model Performance**

|  |  |  |  |
| --- | --- | --- | --- |
| **Metric (test set)** | **Basic Logistic** | **Weighted Logistic** | **LDA** |
| Mis-class rate | 0.163 | 0.213 | 0.195 |
| Sensitivity (Serious) | 0.34 | 0.64 | 0.19 |
| Specificity (Non-Serious) | 0.982 | 0.831 | 0.987 |
| Balanced accuracy | 0.661 | 0.736 | 0.589 |
| AUC | 0.797 | 0.798 | 0.775 |

* The basic logistic model excels at overall accuracy and nearly eliminates false alarms, but class-imbalance bias limits its ability to detect "Serious" vent flow events
* In comparison, the Weighted logistic underperforms ($\approx$ 5%) in accuracy but doubles recall, delivering the best-balanced accuracy and AUC while maintaining acceptable specificity among our models. In more practical/applied settings, if the cost of missing a "Serious"(false negative) event is quite high, this model might be preferred.
* The LDA model performs worst overall mainly due to the equal-covariance and normality assumptions being violated, causing severe under-detection of the minority class.

Actionable insights/Operational context applications:

* The weighted logistic regression could be used as the primary classifier. It mitigates class-imbalance bias, achieves the highest balanced accuracy (0.74) and AUC (0.80), and identifies nearly two-thirds of Serious wells while keeping false alarms manageable.
* The basic logistic may be retained as a fallback when specificity must exceed 0.97.

###### **Model interpretations**

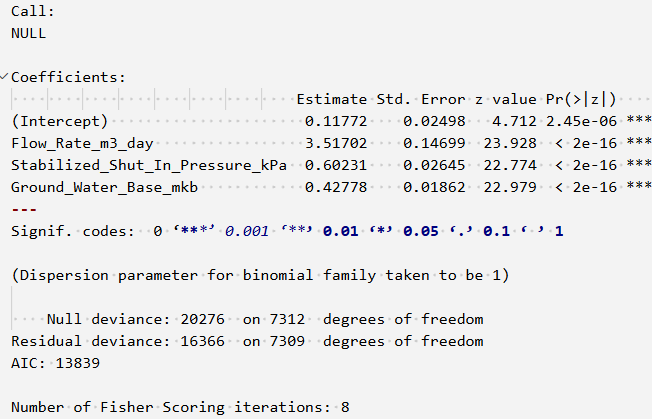


Figure 24: Weighted Logistic regression model summary

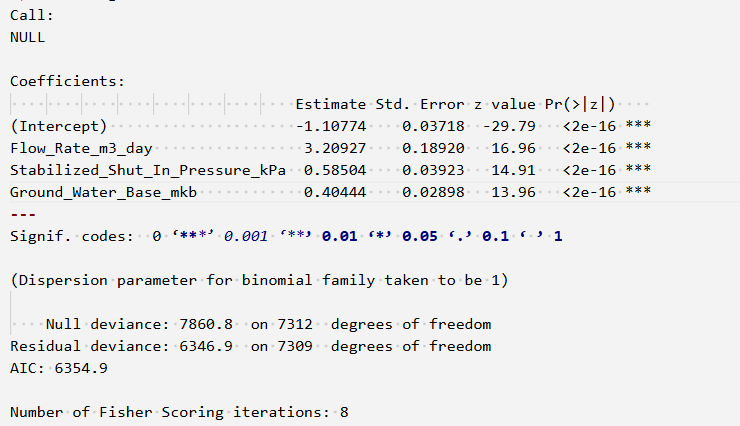


Figure 25: Basic logistic regression model summary

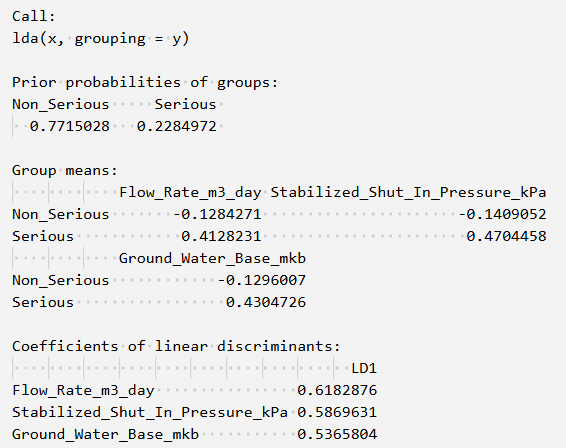


Figure 26: LDA model summary

All methods (LDA, basic logistic, and weighted logistic) converge on the same story that is wells that vent larger gas volumes, originate from higher shut-in pressures, and intercept a deeper ground-water base are most prone to Serious casing-vent events.

"Flow-Rate (m³ day⁻¹)" has the highest discriminant loading (LDA = 0.62) and largest logit coefficient ($\beta ≈ 3.2$). Weighted logit inflates it further ($\beta ≈ 3.5$).

"Stabilised Shut-In Pressure (kPa)" has the second-largest loading and coefficient in every fit. A unit increase multiplies odds by approximately 1.8–2.0.

"Ground-Water-Base Depth (m kb)" has the smallest but still strongly positive loading/coefficient.

The models all point to the same three warning signs of a serious vent-flow problem. First, wells that are blowing off larger volumes of gas each day("Flow-Rate (m³ day⁻¹)") are much more likely to have a big leak simply because more gas is escaping, so the odds of a severe event increase.

Second, wells with higher shut-in pressures have more force pushing the gas upward. That extra pressure can drive gas through cracks or poor-quality cement and turn a small leak into a major one.

Finally, wells which have their ground-water contact lying deeper tend to draw gas from lower, higher-pressure zones. That makes the leak path longer, harder to seal and therefore riskier.

Together these three features mark wells most prone to serious surface-casing vent-flow events.

##### **Future Work**

###### **Final Logistic Regression Model Assumptions and Diagnostics**

* Linearity assumption: The logit plot for the fitted base logistic model still shows noticeable curvature. "Flow Rate" shows some non-linearity in the form of an upward bend. The downward bends in "Ground-Water Base" suggests a concave relationship, though "Shut-In Pressure" has a milder curvature. The linearity assumption is still not fully met. Further work could be done in addressing the non-linearity issue to potentially help improve model performance and generalizability.

### **QUADRATIC DISCRIMINATIVE ANALYSIS (QDA)**

#### Splitting Data into Training and Testing datasets

To ensure reliable and unbiased evaluation of the classification model, the distribution of the target variable Classification—which contains the classes Serious and Non-Serious—was preserved in both the training and testing datasets through stratified sampling. This approach helped prevent bias toward the majority class and ensured that the model’s ability to distinguish between Serious and Non-Serious cases was fairly evaluated.

A screenshot of a computer code

AI-generated content may be incorrect.

A screenshot of a computer code

AI-generated content may be incorrect.

Figure 27: Splitting Data

#### Building QDA Model

First, the QDA model was built using the numerical variables **flow\_rate\_m3\_day**, **stabilized\_shut\_in\_pressure\_kPa**, **source\_depth\_mkb**, and **groundwater\_base\_mkb**. The model was fitted on the training data to classify events based on their regulatory classification.

A screenshot of a computer program

AI-generated content may be incorrect.

Figure 28: Building QDA Model

The summary of the model is given below.

A screenshot of a computer program

AI-generated content may be incorrect.A black text on a white background

AI-generated content may be incorrect.

Figure 29: Model Summary

**Interpretations:**

About 74.3% of vent flow and gas migration events are classified as Non Serious, while 25.7% are Serious. The model accounts for this imbalance by expecting Non Serious events to be more common when making predictions.

The Flow Rate is significantly higher in Serious cases (approximately 404 m3/day versus 20.5 m3/day), indicating that high flow rate is associated with more severe incidents.

Shut-in Pressure is also more than twice as high in Serious cases (around 930.5 kPa compared to 378 kPa), serving as another strong indicator of severity.

Both Source Depth and Ground Water Base are greater in Serious cases, suggesting that more serious issues tend to occur in deeper wells.

**Prediction on Test Data**

Predictions were generated on the test dataset using the QDA model trained on the numerical variables.

A screenshot of a computer code

AI-generated content may be incorrect.

Figure 30: QDA Prediction

The performance of the QDA model was assessed using a confusion matrix. The model correctly classified 1444 Non-Serious events and 140 Serious events but misclassified 379 Non-Serious events as Serious, and 59 Serious events as Non-Serious.

A close-up of a white background

AI-generated content may be incorrect.

Figure 31: QDA - Misclassification rate

The misclassification rate was found to be approximately 0.2166 (21.66%). This indicates that about 21.66% of the cases in the test data were incorrectly classified by the QDA model.

An accuracy of approximately 78.34% was therefore achieved by the model.

##### Confusion Matrix

A confusion matrix was created to evaluate the performance of the classification model by comparing the predicted and actual class labels. It was used to compute key evaluation metrics such as sensitivity, specificity, precision, and balanced accuracy.

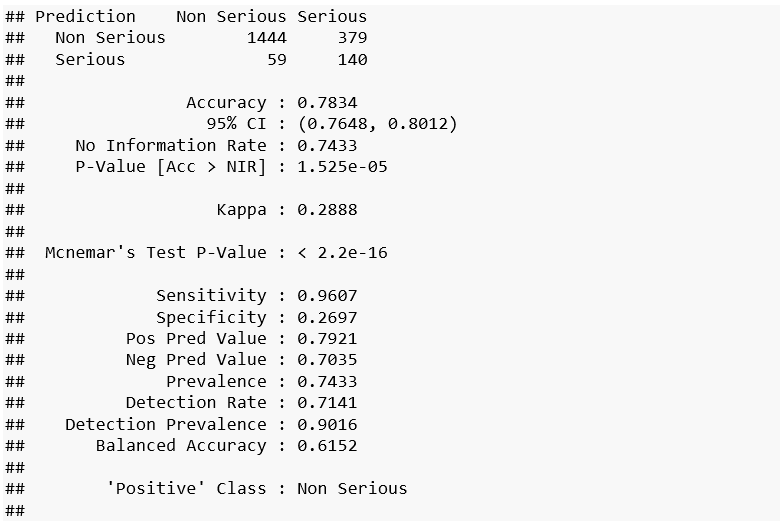


Figure 32: QDA- Confusion Matrix

The following evaluation metrics were obtained to assess the model's performance.

A sensitivity (recall) of 0.9621 indicates that 96.2% of Non-Serious cases were correctly identified.

The specificity was found to be 0.2703, meaning only 27% of Serious cases were correctly classified.

Precision (positive predictive value) was calculated as 0.7926, indicating that when the model predicted a case as Non-Serious, it was correct approximately 79% of the time.

The negative predictive value was 0.7107, suggesting that the model correctly identified Serious cases around 71% of the time.

The balanced accuracy was computed as 0.6162, reflecting the average of sensitivity and specificity, which is particularly relevant in the context of imbalanced classes.

A very low p-value (< 2.2e-16) was obtained, indicating a statistically significant difference in the misclassification patterns. This suggests that the model’s errors are not symmetric, and the model tends to misclassify one class more than the other.

The model was observed to perform well in identifying Non-Serious cases but demonstrated limited ability to detect Serious cases, as evidenced by the low specificity. A tendency to favor the majority class (Non-Serious) was noted (due to the class imbalance).Although an overall accuracy of 78% was achieved, the low specificity and kappa value highlighted challenges in accurately identifying the minority class.

##### **Cross Validation**

To ensure that the performance of the QDA model was not dependent on a single train-test split, stratified 10-fold cross-validation was applied. This approach allowed the model to be trained and validated on multiple balanced subsets of the data, providing a more reliable and stable estimate of its generalization performance.

A screenshot of a computer program

AI-generated content may be incorrect.

Figure 33: QDA - Cross Validation

**Model Performance Evaluation**

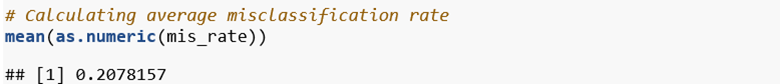


Figure 34: Model Performance - QDA

An average misclassification rate of 0.2078 was observed across the cross-validation folds. This indicates that approximately 79.2% of the events were correctly classified, while the remaining 20.8% were misclassified, suggesting a moderately good performance of the QDA model in distinguishing between Serious and Non-Serious events.

The misclassification rate decreased from 21.66% to 20.78% after cross-validation, indicating a slight improvement in the model’s generalizability.

**Confusion Matrix**

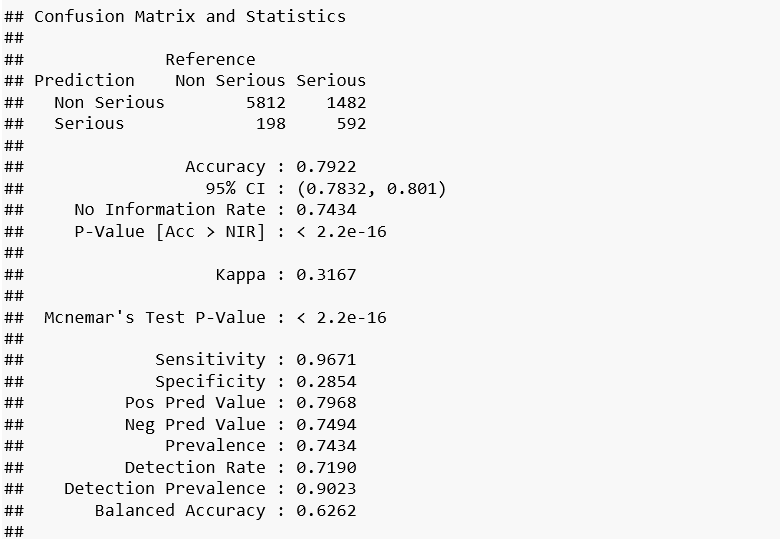


Figure 35: Confusion Matrix- QDA Cross validation

The current model shows a sensitivity (recall) of 0.9671, indicating a very high ability to correctly identify Non-Serious cases, slightly improving upon the previous value of 0.9607. However, the specificity remains low at 0.2854, meaning only about 28.5% of Serious cases are correctly classified, which is a minor increase from the previous 26.97%. The precision (positive predictive value) is 0.7968, showing that when the model predicts Non-Serious, it is accurate approximately 80% of the time, consistent with the earlier 79.21%. The negative predictive value improved slightly to 0.7494, meaning predictions of Serious cases are correct around 75% of the time, up from 70.35%. Finally, the balanced accuracy increased to 0.6262, reflecting a better average performance across both classes compared to the previous 61.52%, indicating modest improvement in handling class imbalance.

#### **QDA – Checking Assumptions**

##### 1. Multivariate Normality:

Within each class (e.g., Serious, Non Serious), the predictor variables are assumed to follow a multivariate normal distribution. X | Y = k ~ N (μ\_k, Σ\_k)

Multivariate Normality Test (Mardia’s Test) was used to check multivariate normality assumption.The library ‘QuantPsyc’ was used.

Null hypothesis (H₀): The predictor variables follow a multivariate normal distribution within each class.

Alternative hypothesis (H₁): The predictor variables do not follow a multivariate normal distribution within each class.



A screenshot of a computer program

AI-generated content may be incorrect.

A screenshot of a test

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 36 Multivariate Normality Test Results

The test was performed separately on the Serious and Non Serious groups. In both groups, the p-values for skewness and kurtosis were 0, which is less than the significance level of 0.05. This means we reject the null hypothesis that the data are multivariate normal.

Therefore, the predictor variables are not normally distributed within either group, violating the normality assumption of QDA. As a result, the reliability of the QDA classification results may be compromised.

The Mahalanobis QQ plot was used to check if the predictor variables follow a multivariate normal distribution.

A screenshot of a computer code

AI-generated content may be incorrect.

A computer screen shot of a computer code

AI-generated content may be incorrect.





A graph with a line and a line

AI-generated content may be incorrect.

Figure 37 QQ Plot - Serious Class

A graph of a graph

AI-generated content may be incorrect.

Figure 38 QQ Plot - Non Serious Class

The Mahalanobis QQ plots for both classes showed significant deviation from the reference line on the right side. This suggests the presence of outliers or heavy tails, indicating that the predictor variables are not multivariate normal, especially in the upper range. As a result, the normality assumption of QDA is violated.

##### 2. Class-Specific Covariance Matrices (Σ\_k):

Each class has its own covariance matrix (Σ\_k) which allows QDA to model more flexible (nonlinear) decision boundaries compared to LDA.

To examine the structure and spread of the predictor variables within each class, covariance matrices were computed and separately for the Serious and Non-Serious groups.

A close-up of a computer code

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A white background with black text

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 39 Covariance Matrix - Serious and Non Serious Classes

Since the covariance matrices vary significantly between the Serious and Non-Serious groups, QDA’s assumption that each class has its own covariance matrix is achieved.

**Covariance Matrices - Visualization using Heat Map**

Due to large differences in the magnitude of covariance values between the two groups, a signed log transformation was applied to normalize the scale and allow for meaningful visual comparison.

The transformed covariance matrices were then visualized as heatmaps using ggplot2. This approach helped to highlight both the direction and strength of relationships between variables within each class while ensuring they were shown on a consistent scale.

A screen shot of a computer code

AI-generated content may be incorrect.

A computer screen shot of a program

AI-generated content may be incorrect.

A computer screen shot of a computer code

AI-generated content may be incorrect.

A graph with blue squares

AI-generated content may be incorrect.

Figure 40 Covariance Matrix - Non Serious Class

A graph with blue squares

AI-generated content may be incorrect.

Figure 41 Covariance matrix - Serious Class

The significant difference observed between the covariance matrices of the Serious and Non*-*Serious classes supports the key assumption of QDA that each class has its own distinct covariance structure.

##### 3. Independence of Observations:

All observations are assumed to be independent of each other and drawn from the same distribution within each class.

### **QDA Model – (Flow Substance Included)**

A screenshot of a computer

AI-generated content may be incorrect.

Before building the QDA model, the categorical variables Type and Flow\_Substance were examined for class balance and category frequency. Since QDA relies on estimating class-conditional covariance matrices, it is essential that categorical predictors have sufficient and reasonably balanced representation across both classes to avoid instability and overfitting.

Some categories in the Flow\_Substance variable, such as Condensate and Saline, were very rare. These categories appeared mostly in the Serious class and had too few cases for stable modeling with QDA.

To solve this, all rare categories were grouped into a new category called “Other”. This new group had 43 out of 44 cases in the Serious class. So, the class-related pattern was still kept.

This change made the model simpler and helped improve the reliability of QDA.

A screenshot of a computer program

AI-generated content may be incorrect.A screenshot of a computer

AI-generated content may be incorrect.

Since the dataset has been cleaned and certain categories have been combined, it is necessary for unused factor levels to be removed. This ensures that only the relevant, currently observed levels are retained in the modified categorical variables. By doing so, any confusion during modeling and interpretation caused by obsolete categories is avoided.

A screenshot of a computer program

AI-generated content may be incorrect.

The modified dataset was then split into training and test data.

A screenshot of a computer code

AI-generated content may be incorrect.

A screenshot of a computer program

AI-generated content may be incorrect.A screenshot of a computer program

AI-generated content may be incorrect.A screenshot of a computer

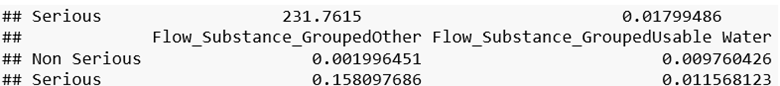
AI-generated content may be incorrect.

Figure 42 QDA Model Summary

A proportion of 74.34% of the observations in the training data was classified as Non-Serious, while 25.66% was classified as Serious.

Larger group means were observed in the serious class for all four numerical predictors, indicating that higher values leads to Serious events.

For the Flow\_Substance\_Grouped variable:

1. The “Other” category has a much higher mean in the Serious class (15.8%) than in the Non-Serious class (0.2%), indicating it is strongly associated with Serious cases.
2. This shows that all the flow substances merged with the ‘Other’ category are related to Serious events.
3. Other grouped categories (like “Not Converted” and “Usable Water”) show only small differences between the two classes.

**Making predictions on test data**

A screenshot of a computer code

AI-generated content may be incorrect.

A total of 1445 observations were correctly predicted as Non-Serious.

191 observations were correctly predicted as Serious.

57 serious observations were incorrectly predicted as Non Serious (false negatives).

327 non-serious observations were incorrectly predicted as Serious (false positives).

**Misclassification Rate**

A black and orange text

AI-generated content may be incorrect.

Figure 43 Misclassification rate - QDA Model

A misclassification rate of approximately 19.01% was observed. This means that 19.01% of the test samples were misclassified by the QDA model, while the remaining 80.99% were correctly classified.

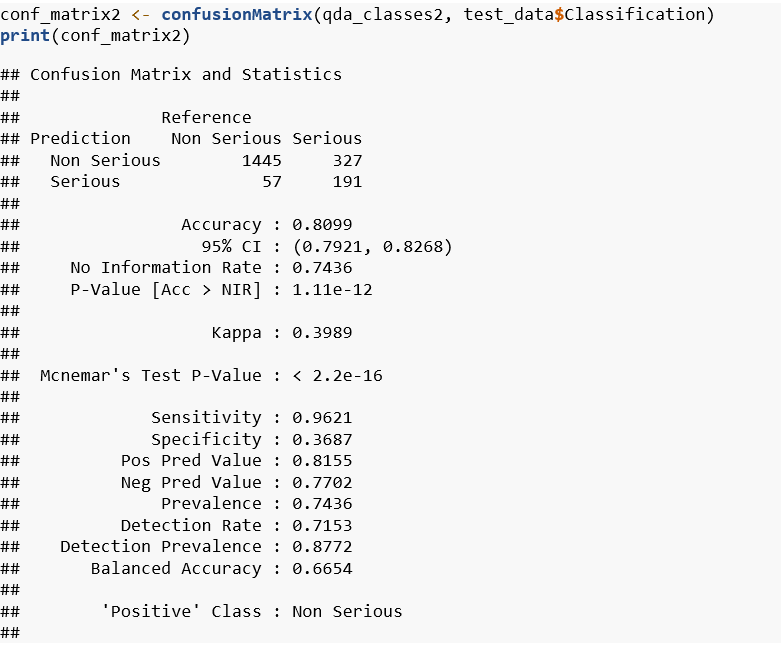


Figure 44 Confusion Matrix - QDA Model

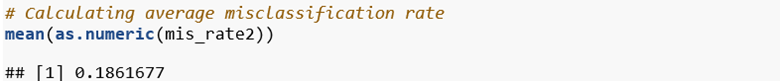
All performance metrics - Sensitivity,Specificity,Precision, Pos and Neg pred values show an improvement compared to the model using numerical predictors only.

1. Prevalence (0.7436): In the dataset, 74.36% of the actual cases belong to the “Non Serious” class, indicating a class imbalance toward this category.
2. Detection Rate (0.7153): The model correctly identified 71.53% of all samples as “Non Serious”, which reflects its ability to detect the positive class.
3. Detection Prevalence (0.8772): The model predicted 87.72% of the cases as “Non Serious”, showing that it tends to favor this class in its predictions.

#### Cross Validation – 10-fold Stratified Sampling

A screenshot of a computer program

AI-generated content may be incorrect.



A misclassification rate of 19.01% was observed prior to cross-validation, while a slightly lower rate of 18.62% was obtained after performing cross-validation. This decrease suggests that the model’s predictive performance was slightly improved when evaluated more robustly through cross-validation.

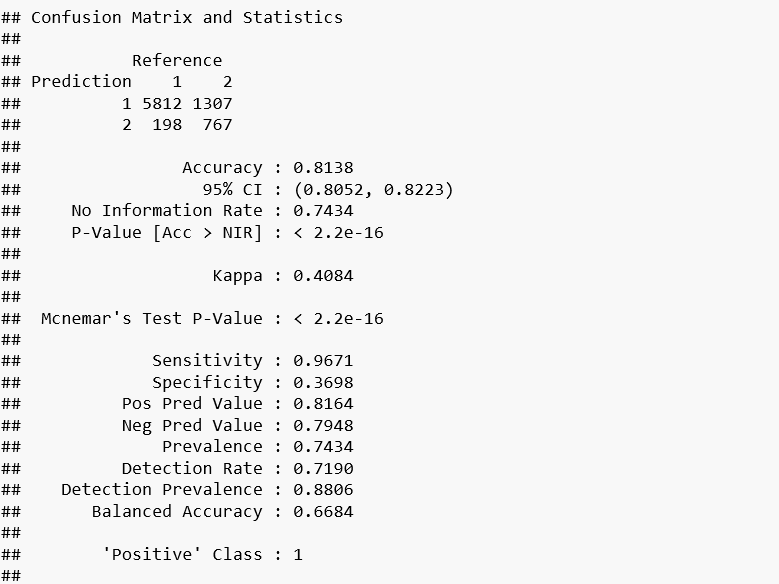


Figure 45 Confusion Matrix -QDA Model

Following cross-validation, the model demonstrated slight improvements in overall accuracy, agreement, and predictive performance. Sensitivity and precision remained high, while balanced accuracy showed modest gains. McNemar’s test continued to indicate a significant difference in misclassification patterns. These results suggest more robust and reliable model performance after validation.

### **CLASSIFICATION TREE**

#### Classification Tree with Stratified Train-Test Split

A classification tree was constructed to predict vent-flow severity, using a stratified sampling approach to split the data. This method ensured that the ratio of "Serious" to "Non-Serious" cases remained consistent in both the training and test sets, addressing the substantial class imbalance (6,010 Non-Serious, 2,074 Serious cases). The training set contained 4,507 Non-Serious and 1,555 Serious cases, while the test set had 1,503 Non-Serious and 519 Serious cases, preserving the original class distribution.

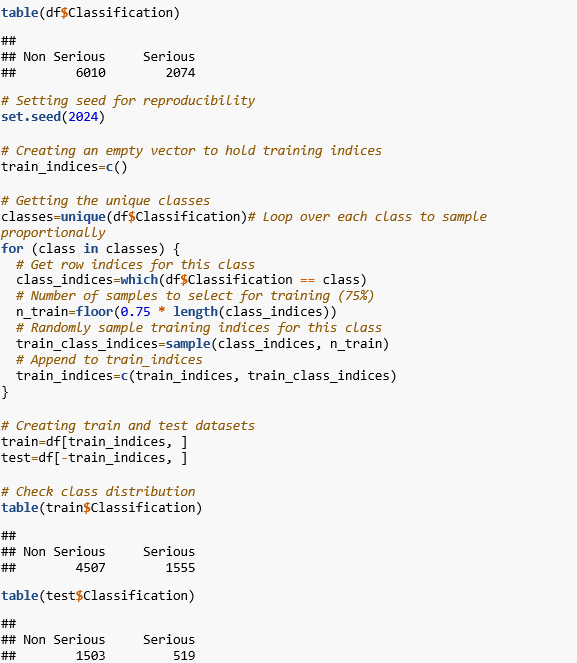


Figure 46: Stratified Train-Test Split

The decision tree model was built using the following predictors: Flow Rate (m³/day), Stabilized Shut-In Pressure (kPa), Source Depth (mKB), Ground Water Base (mKB), Flow Substance. The tree helps us understand which conditions are most predictive of serious vent-flow events in Alberta’s oil and gas wells.

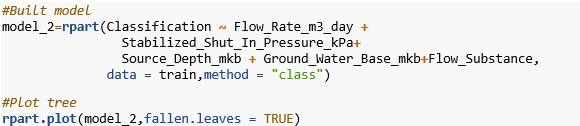


Figure 47: Classification Tree - model building

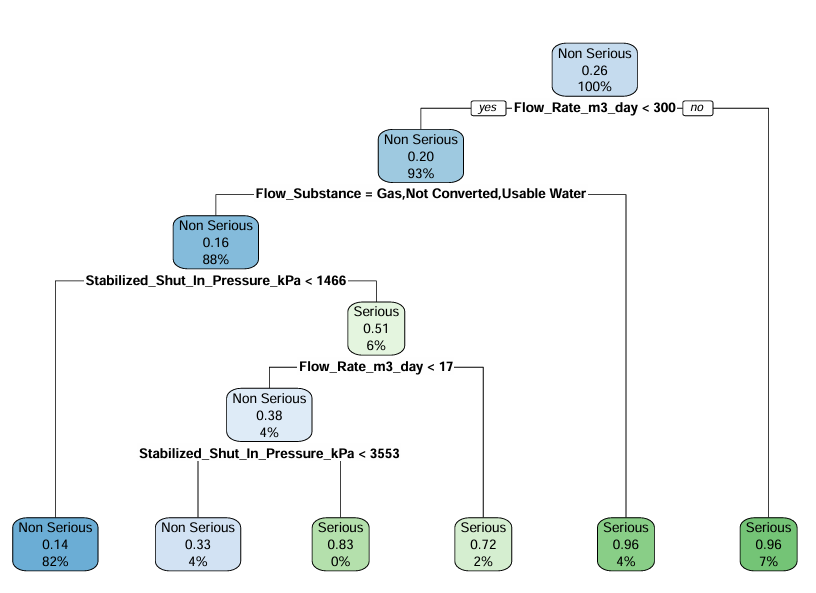


Figure 48: Tree

#### Key Findings from the stratified classification tree:

The key features used in the tree are Flow Rate (m³/day), Stabilized Shut-In Pressure (kPa), Flow Substance.

The tree works by sequentially splitting the data based on the most informative features, creating a set of decision rules that lead to a final classification at each terminal node.

* Flow Rate is the Primary Split Criterion  
   The first and most important decision point in the tree is the flow rate:
  + If Flow\_Rate\_m3\_day greater than 300, the event is almost always classified as Serious.
  + If Flow\_Rate\_m3\_day less than 300, further conditions are assessed to determine severity.
* Low Flow Rate + Flow Substance Type => Non-Serious  
   Among events with lower flow rates (300 m³/day), the type of flow substance plays a major role:
  + If the substance is Gas, Not Converted, or Usable Water, the event is highly likely to be Non-Serious.
* High Shut-In Pressure Can Still Lead to Serious Events  
   If the flow substance does not belong to the above group, then shut-in pressure becomes critical:

When Stabilized\_Shut\_In\_Pressure\_kPa greater than 1466, further splits are made based on flow rate and pressure.

* + - If Flow\_Rate\_m3\_day greater than 17, the event is Serious.
    - Even with lower flow rates (less than17), higher shut-in pressures (greater than 3553) still lead to Serious classification.
* Low Pressure Events Are Generally Non-Serious  
   If both flow rate is low and pressure is below 1466 kPa, the tree consistently classifies events as Non-Serious.

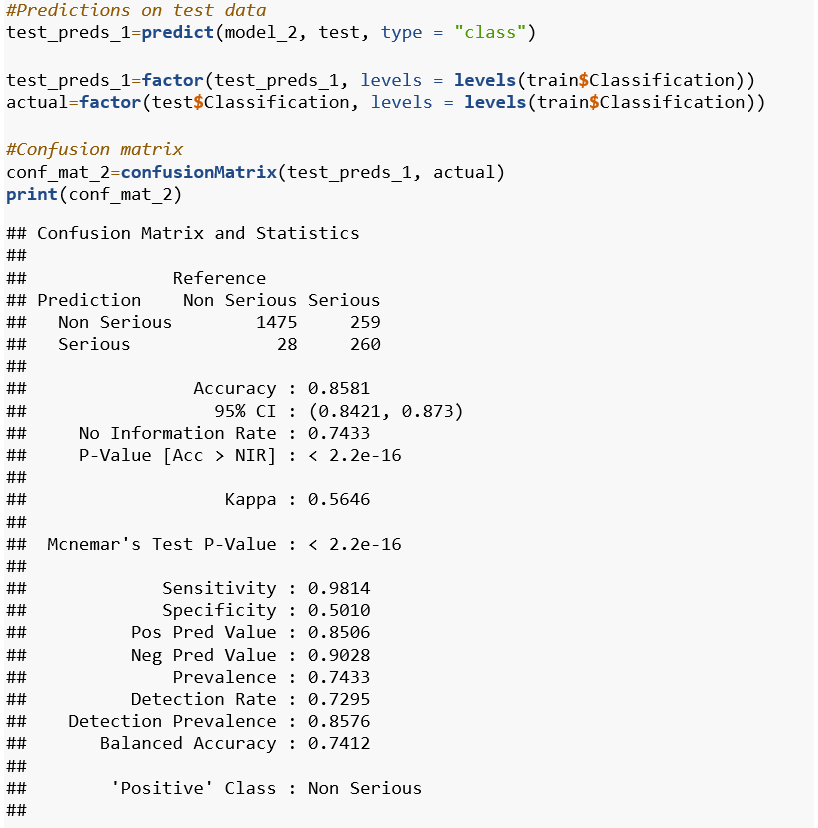


Figure 49: Tree Confusion Matrix

The model correctly identified the majority of "Non-Serious" cases (1,475 out of 1,503), but was less accurate for "Serious" cases (260 out of 519). The higher sensitivity for "Non-Serious" reflects the class imbalance, as there are more "Non-Serious" examples in the dataset. The moderate specificity indicates that some "Serious" cases are misclassified as "Non-Serious". And the accuracy is 0.8581.

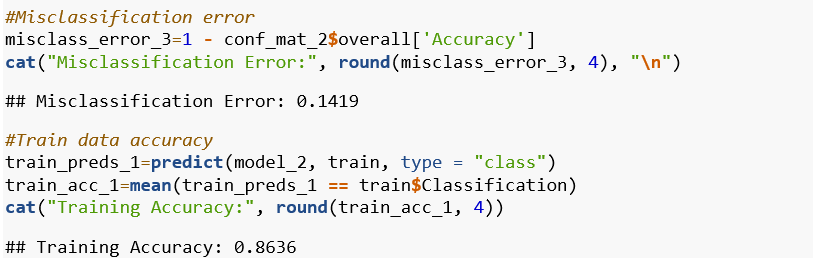


Figure 50: Metric Calculation

The misclassification error on the test set was 0.1419 (or 14.19%). This means that about 14.2 out of every 100 predictions made by the model on new, unseen data were incorrect.

The small difference between training and test accuracy (0.8636 vs. 0.8581) suggests that the model generalizes well and is not overfitting.

#### **Classification Tree with Stratified K-Fold Cross-Validation**

To further ensure balanced evaluation, a stratified K-Fold cross-validation approach was used. This technique maintained the proportion of "Serious" and "Non-Serious" cases in each fold, providing a fair assessment despite class imbalance.



Figure 51: K-fold Classification tree building

The data is split into 10 parts (folds), each with about the same ratio of "Non-Serious" and "Serious" cases. For each fold, the model is trained on 9 parts and tested on the 1 part left out. This is repeated 10 times, so every data point is used for both training and testing. Results are averaged to give a reliable estimate of model performance, reducing bias from any single split.



Figure 52: K-fold tree

#### **Key Findings from the Stratified K-Fold Cross Validation tree:**

The key features used in the tree are Flow Rate (m³/day), Stabilized Shut-In Pressure (kPa), Flow Substance and Source\_Depth\_mkb.

* Flow Rate is the Most Critical Predictor

If flow rate is greater than 299 m³/day, the event is classified as Serious with 96%

Confidence.

High flow rates alone are a strong indicator of Serious vent-flow events.

* Flow Substance Type Helps Identify Non-Serious Events
  + For events with low flow rate (<299 m³/day), the tree checks the flow substance.
  + If it is Gas, Not Converted, or Usable Water, the event is likely Non-Serious (88%).

When the flow rate is low and the substance is common or benign, the event is generally Non-Serious.

* Pressure and Depth Reveal Hidden Serious Risks

For flow substances not in the above category, the model assesses two additional factors:

* + Stabilized Shut-In Pressure (kPa)
    - If pressure is <1469, the event is mostly Non-Serious (82%)
    - If pressure is ≥1469, the model uses source depth for a finer decision.
  + Source Depth (mKB)
    - If depth < 0.55, the event is Serious (66%)
    - If depth ≥ 0.55, the event is Serious (96%)

Even if flow rate is low, high shut-in pressure combined with deeper source zones increases the risk of a Serious event.

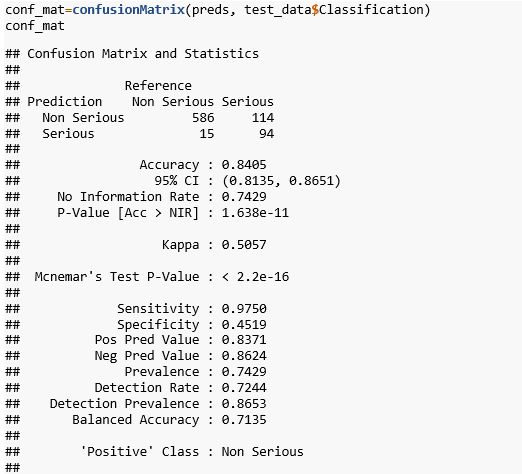


Figure 53: K-fold tree confusion matrix

The model correctly identified the majority of "Non-Serious" cases (586 out of 601) but was less accurate for "Serious" cases (94 out of 208). The high sensitivity for "Non-Serious" reflects the class imbalance, as there are more "Non-Serious" examples in the dataset. The lower specificity indicates that a significant number of "Serious" cases were misclassified as "Non-Serious." This pattern is typical when one class is much larger than the other, leading the model to favor the majority class. Balanced accuracy and Kappa statistics were reported to provide a fairer assessment of performance in the presence of class imbalance.

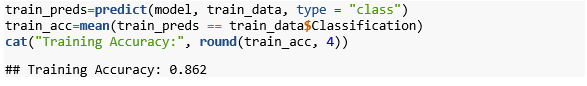


Figure 54: K-fold tree metric calculation

The small difference between training and test accuracy (0.862 vs. 0.858) indicates that the model generalizes well and is not overfitting. On average, the model makes errors on 14.18% of the test samples. Model's reliable performance when predicting new data.

# **CONCLUSION**

Based on the analysis of all four classification models - Logistic Regression, Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), and Classification Tree applied to the Surface Casing Vent Flow (SCVF) and Gas Migration dataset.

1. Which characteristics are best predictors for regulatory classification (Serious, Non-Serious)?

* **Most influential predictors** for distinguishing between "Serious" and "Non-Serious" events are flow rate (m³/day), stabilized shut-in pressure (kPa), source depth (mKB), and groundwater base depth (mKB).
* Among categorical variables, flow substance (with gas being the most common in both serious and non-serious events).

2. Can Linear Discriminant Analysis (LDA) or Quadratic Discriminant Analysis (QDA) effectively classify events based on the available features?

* LDA is **not reliable** for this dataset. Its core assumptions are violated, particularly normality and equal variance-covariance matrices. Using LDA here would likely result in poor accuracy, biased classification boundaries, and unreliable probability estimates, especially for the minority ("Serious") class. LDA's linear boundaries are not suitable for the observed data distribution.
* QDA is **somewhat more appropriate** than LDA for this dataset because it allows each class to have its own covariance structure, accommodating the observed heteroscedasticity (unequal variances). However, the violation of the multivariate normality assumption still limits its effectiveness. While QDA perform better than LDA, its predictions and probability estimates could still be affected by the skewed, heavy-tailed distributions and extreme outliers present in the data.

3. How accurately do our classification models predict the severity of events when evaluated with cross‐validation and confusion matrix?

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Misclassification Rate** | **Accuracy** | **Correctly Identified as "Non-Serious"** | **Correctly Identified as "Serious"** |
| Logistic Regression​  (with class weights)​ | 21.26 %​ | 78.74 %​ | 83.09 %​ | 64.09 %​ |
| LDA​ | 19.53 %​ | 80.47 % ​ | 98.67 % ​ | 19.03 % ​ |
| QDA​ | 18.62%​ | 81.38%​ | 96.71%​ | 36.98 %​ |
| Classification Tree​  (No weights)​ | 14.19%​ | 85.8%​ | 98.14 %​ | 50.1%​ |

* Logistic Regression (with class weights):
  + Achieves the highest recall for "Serious" events (64.09%), aligning with the goal of maximizing identification of serious cases.
  + Has lower overall accuracy (78.74%) and a higher misclassification rate (21.26%) compared to other models.
* Classification Tree (No weights):
  + Delivers the highest overall accuracy (85.8%) and the lowest misclassification rate (14.19%).
  + Shows a bias toward the "Non-Serious" class, with lower recall for "Serious" cases (50.1%), likely due to class imbalance.
* LDA and QDA:
  + Highly effective at identifying "Non-Serious" events (LDA: 98.67%, QDA: 96.71%).
  + Less effective for "Serious" events (LDA: 19.03%, QDA: 36.98%), making them less suitable for maximizing identification of serious cases.

**Model Selection**  
Logistic Regression (with class weights) is the most effective model for maximizing the identification of "Serious" cases, despite its lower overall accuracy and higher misclassification rate. The Classification Tree provides the best overall accuracy but is less effective at identifying "Serious" cases due to class imbalance.

### **Future Improvement**

To address class imbalance and improve model performance across all classes, we can incorporate techniques such as SMOTE (Synthetic Minority Over-sampling Technique), random oversampling, and class weighting. These methods can help ensure that minority classes receive adequate representation during model training, leading to more balanced and reliable predictions.

In this dataset, we currently utilize one categorical predictor and four numerical predictors. For future research or model enhancement, we could explore additional relevant features or measurements that may influence the severity of events. Identifying and incorporating such variables could improve model accuracy and robustness, particularly in predicting high-severity cases.

# **REFERENCES**

1. General data. (2025 May 27). Alberta Energy Regulator. <https://www.aer.ca/data-and-performance-reports/activity-and-data/general-well-data>
2. License [Copyright and Disclaimer | Alberta Energy Regulator](https://www.aer.ca/copyright-and-disclaimer)
3. Schiffner, D., Kecinski, M. & Mohapatra, S. (2021) An updated look at petroleum well leaks, ineffective policies and the social cost of methane in Canada’s largest oil-producing province. Climatic Change, 164(3), p. 60. doi: 10.1007/s10584-021-03044-w.
4. BC Energy Regulator (n.d) Well Emissions. Available at: <https://www.bc-er.ca/data-reports/data-narratives/well-emissions/> (Accessed: 30 May 2025).
5. Zach Bobbitt, How to Perform Multivariate Normality Tests in R <https://www.statology.org/multivariate-normality-test-r/>
6. Ripley, B. & R Project. (2024). *Classification and regression trees* (B. Ripley, Ed.; Version 1.0-44). CRAN. <https://cran.r-project.org/web/packages/tree/tree.pdf>
7. Kamaldeep. (2025, May 1). *How to Improve Class Imbalance using Class Weights in Machine Learning?* Analytics Vidhya. <https://www.analyticsvidhya.com/blog/2020/10/improve-class-imbalance-class-weights/>
8. *SMOTE function - RDocumentation*. (n.d.). <https://www.rdocumentation.org/packages/DMwR/versions/0.4.1/topics/SMOTE>
9. *RPubs*. (n.d.). <https://rpubs.com/>