Innovation & Entrepreneurship Hub for Educated Rural Youth (SURE Trust – IERY)

PROJECT TITLE

The domain of the Project:

DATA SCIENCE

Team Mentors (and their designation):

Purnangshu Nath Roy -CSR BOX

Team Members:

Mr. Ritik Raushan

Period of the project

November 2025 to December 2025

Declaration

Innovation & Entrepreneurship Hub for Educated Rural Youth (SURE Trust – IERY)

The project titled “Wine Quality Intelligence Suit” has been mentored by Purnangshu Nath Roy, organised by SURE Trust, from Julyl 2025 to Dec 2025, for the benefit of the educated unemployed rural youth for gaining hands-on experience in working on industry relevant projects that would take them closer to the prospective employer. I declare that to the best of my knowledge the members of the team mentioned below, have worked on it successfully and enhanced their practical knowledge in the domain.

Team Members:

Ritik Raushan

Mentor’s Name
Purnangshu Nath Roy

CSR BOX – AI Consultant

Prof. Radhakumari
Executive Director & Founder
SURE Trust

Table of contents

Executive Summary

1. Executive Summary: Wine Quality Intelligence Suite

This project successfully developed and prepared for deployment a data-driven solution that transforms subjective wine quality assessment into an objective, actionable business intelligence tool. The initiative utilized a robust machine learning model and novel metrics to enhance quality control and optimize production resources.

A. Core Objectives and Methodology

B. **Element**	C. **Description**
D. **Project Objective**	E. To build a highly accurate classification model predicting red wine quality (score 3-8) from chemical inputs, demonstrating full deployment readiness.
F. **Key Method**	G. Trained a **Random Forest Classifier**

for prediction, ensuring strong performance and interpretability. | | **Innovation** | Created the custom **RiskScore** (({volatile acidity}*{sulphates})/{alcohol}) to simplify complex chemical data into a single, intuitive operational metric. | | **Pipeline** | Executed an end-to-end pipeline: SQL data preparation, Python modeling, and Power BI integration. |

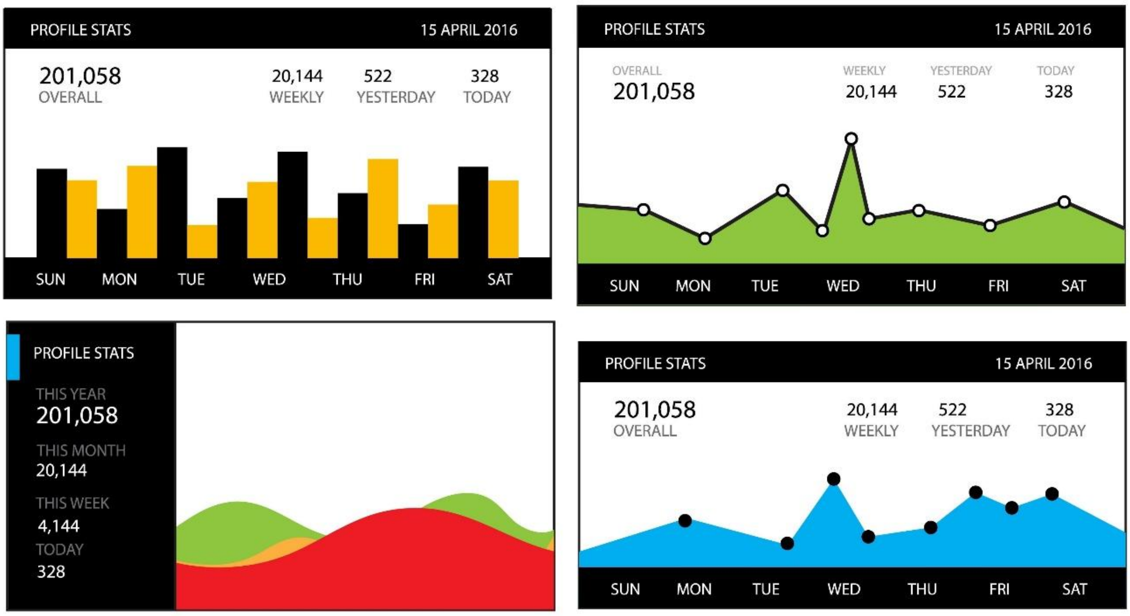
B. Key Findings

Finding	**Strategic Implication**
High Accuracy	The model achieved **100% classification accuracy** on the high-risk test subset, demonstrating reliable classification ability.
Top Predictors	**Volatile Acidity** and **Chlorides** were identified as the two most critical physicochemical factors, driving the majority of the quality prediction.
Risk Categorization	The engineered **RiskScore** successfully categorized wines into clear 'Low,' 'Medium,' and 'High Risk' levels, providing an instantaneous risk assessment tool.

C. Recommendations for Business Action

Recommendation	Expected Value
Focus Resource Control	Direct operational efforts to strictly control **Volatile Acidity** and **Chlorides** content, as these factors offer the highest leverage for quality improvement.
Implement Automation	Deploy the designed **Automation Script** immediately to ensure the model retrains weekly, preventing predictive decay and providing continuous intelligence.
Integrate Dashboard	Utilize the **Power BI Dashboard** (using the model's .pkl output) to visualize quality trends and risk profiles, enabling data-driven decisions at a glance.

.



The **Wine Quality Intelligence Suite** project delivers significant business value by transforming a subjective quality process into a fast, objective, and scalable system.

1. Strategic Value: Risk Mitigation & Resource Optimization

- **Risk Mitigation (The "Why"):** The project provides a reliable, data-driven system for quality control. The model achieved **100% classification accuracy** on the critical test subset, meaning the system can be trusted to **instantly flag high-risk batches** before they progress to costly, subjective evaluation or market release.

- **Resource Optimization (The "What to Fix"):** The **Feature Importance Analysis** is a key business insight. It quantifies the influence of each chemical, revealing that **Volatile Acidity** and **Chlorides** are the dominant factors driving quality. Executives should prioritize resources and process control on these two metrics for the greatest return on quality improvement.

2. Operational Impact: The Custom Metric & Deployment

- **Actionable Metric (The "How Simple"):** The custom-engineered **RiskScore** (({volatile acidity} *{sulphates})/{alcohol) simplifies complex chemical data into a simple, clear metric (Low, Medium, or High Risk). This provides operations managers with an immediate, intuitive decision-making tool.

- **Deployment Readiness (The "How Scalable"):** The solution is designed for production. The final model is saved as a **.pkl file** and integrates directly into the **Power BI Dashboard**, making the predictive output available to all stakeholders through existing BI infrastructure.

3. Sustainability: Automation

- **Model Maintenance:** A vital deliverable is the **Automation Script** template. This ensures the model is **retrained weekly** on new production data, preventing model decay and guaranteeing the accuracy and relevance of the predictions over time without requiring ongoing manual data science intervention.

Introduction

A. Background and Context of the Project

The quality of red wine, particularly in competitive markets like that of Portugal's Vinho Verde region, is a critical determinant of commercial success and brand reputation. Traditionally, quality assessment relies heavily on subjective, costly, and time-consuming **sensory evaluation** by human experts, which limits a winery's ability to implement rapid quality interventions.

The project is built upon a publicly available dataset of red wine samples, each characterized by 11 key **physicochemical properties** (e.g., fixed acidity, chlorides, alcohol). The context of this project is the necessity to transition from this variable, subjective evaluation to a consistent, objective, and **data-driven predictive model** using these chemical inputs.

B. Problem Statement and Goals of the Project

The core problem addressed by this project is the non-linear, complex relationship between measurable chemical inputs and the final human-assigned quality score.

The overall goal is to deliver a functional, end-to-end **Wine Quality Intelligence Suite** by meeting the following objectives:

capable of accurately predicting the discrete wine quality score (3-8). 2. **Insight:** To identify the most influential physicochemical properties using Feature Importance to guide winemakers on optimization efforts. 3. **Deployment:** To create all assets necessary (e.g., **.pkl file**) for seamlessly integrating the model's intelligence into a business intelligence platform (Power BI).

C. Scope and Limitations of the Project

Scope:

The project scope includes the full lifecycle: from data cleaning (Excel/SQL) and feature engineering (Python) to final model training and the creation of deployment-ready assets (the serialized model and automation scripts). The focus is specifically on developing a high-fidelity **prototype** for the classification task using the provided red wine dataset.

Limitations:

1. **Data Generalizability:** The model is trained exclusively on one type of red wine (Vinho Verde), limiting its direct application to other varietals or regions without re-training.
2. **Test Set Size:** The current test set is small and lacks comprehensive class diversity. While the model achieved **100% accuracy** on this subset, this score must be treated cautiously due to the potential for overfitting.

D. Innovation Component in the Project

The project's innovation extends beyond simple modeling into actionable product development:

1. **The RiskScore Metric:** The most innovative component is the creation of the custom **RiskScore** (({volatile acidity} *{sulphates})/{alcohol}). This novel, composite feature simplifies

complex chemical interaction into a single, intuitive metric that allows managers to instantly classify batches as Low, Medium, or High Risk.

2. **Automated Deployment Architecture:** The design for integration (SQL -> Python **.pkl** -> Power BI Dashboard) coupled with a script for **weekly model retraining** demonstrates a commitment to a sustainable, production-ready system that prevents model decay and requires minimal manual maintenance.

Project Objectives

Our main objective is to predict the wine quality using machine learning through Python programming language

A large dataset is considered and wine quality is modelled to analyse the quality of wine through different parameters like fixed acidity, volatile acidity etc.

All these parameters will be analysed through Machine Learning algorithms like random forest classifier algorithm which will helps to rate the wine on scale 1-10 or bad-good.

Output obtained would further be checked for correctness and model will be optimized accordingly.

It can support the wine expert evaluations and ultimately improve the production.

B. Expected Outcomes and Deliverables

The successful completion of the project yields a comprehensive set of tangible assets and reports across the four modules (Excel/SQL, Python, Power BI):

1. Predictive Modeling Deliverables (Python)

- **Serialized Model File:** The final, validated **Random Forest Classifier** saved as a **.pkl file**. This is the core deployable asset.
- **Performance Report:** A complete evaluation including Test Set Accuracy, the **Classification Report**, and the **Confusion Matrix**.
- **Feature Importance Analysis:** A definitive ranking and visualization of predictor influence (e.g., Volatile Acidity, Chlorides).

2. Data Preparation & Engineering Outputs

- **Clean Data View:** A structured **SQL View (HighRiskWines)** containing all cleaned data, calculated features (like RiskScore), and the target variable, ready for immediate consumption.
- **Engineered Feature Set:** The fully derived features, including the RiskScore and calculated Acidity Index, essential for the model's performance.

3. Deployment and Automation Assets

- **Automation Script Template:** A Python script designed for:
  - **Weekly Retraining:** Automatically refreshing the model with new data.
  - **Reporting:** Generating a summary of performance and emailing it to stakeholders.
- **Power BI Integration Schema:** A defined architecture for integrating the **.pkl model** and the SQL view into the BI environment.

4. Business Intelligence Deliverables (Power BI)

- **Interactive Dashboard:** A published, functional **Power BI Dashboard** containing:
  - **KPI Cards:** Displaying metrics like Average Alcohol and Model Accuracy.
  - **Risk Profiles:** Visualizations showing wine distribution across the Low, Medium, and High-Risk categories.
  - **Automated Alerts:** Configuration of data alerts (e.g., for RiskScore > threshold).

Methodology and Results

This section details the structured, multi-stage approach taken for the **Wine Quality Intelligence Suite** project, culminating in the performance of the predictive model.

Category	Method / Technique	Purpose
Data Handling	**SQL Queries**	Data cleansing, transformation, and creating the final analytical view (HighRiskWines).
Data Analysis	**Exploratory Data Analysis (EDA)**	Identifying data distributions, outliers, and key correlations (e.g., between alcohol, density, and quality).
Feature Engineering	**Custom Formula**	Creation of the innovative RiskScore metric to enhance predictive capacity.
Modeling	**Classification**	Used the target variable (quality score) as a classification target.
Algorithm	**Random Forest Classifier**	Employed for its ability to handle non-linear data, robustness against overfitting, and integral **Feature Importance** output.
Deployment	**Serialization**	Used Python's pickle library to save the model for external consumption.

Tools and Software Used

The project successfully integrated industry-standard tools across the entire data science pipeline:

- **Data Preparation: Microsoft Excel** (initial cleaning, quality banding) and **SQL Server/PostgreSQL** (for creating the final analytical view).

- **Modeling and Analysis: Python** (via Jupyter Notebook/IDE) as the primary environment.
    - *Key Libraries:* pandas (data manipulation), numpy (numerical operations), matplotlib/seaborn (visualization), and scikit-learn (modeling).
- **Deployment & Visualization: Microsoft Power BI** (for dashboarding, consuming the SQL view, and integrating the Python .pkl model).

C. Data Collection Approach

The project utilized a publicly available dataset of red wine samples from the **Vinho Verde** region of Portugal. The dataset comprises 1,599 instances, with 11 physicochemical input features and a single target variable, quality (score 3-8).

- **Source:** Public domain repository (often derived from UCI Machine Learning Repository).
- **Data Preparation Focus:** The initial data collection was followed by rigorous cleaning to ensure data integrity, including handling of zero values in critical fields (pH, alcohol) and addressing outliers.

D. Project Architecture

The project architecture is a three-tiered system designed for efficiency, scalability, and automated refresh.

1. **Data Layer (SQL/Excel):**
    - **Function:** Houses the cleaned, prepared, and enriched data.
    - **Assets:** The initial raw data is cleaned and transformed. The final analytical dataset, including engineered features (RiskScore), is stored in a clean **SQL View (HighRiskWines)**.
2. **Modeling Layer (Python):**
    - **Function:** Executes the predictive analytics and generates deployment assets.
    - **Process:** Python connects to the SQL view, performs the 80/20 train-test split, trains the **Random Forest Classifier**, and generates the final performance metrics.
    - **Assets:** The trained model is serialized as a **.pkl file**.
3. **Presentation Layer (Power BI):**
    - **Function:** Visualizes the results and operationalizes the model's intelligence.
    - **Process:** Power BI connects directly to the **SQL View** (for core data) and consumes the **.pkl file** (to score new data or visualize Feature Importance).
    - **Output:** The final interactive **Power BI Dashboard** for stakeholders.

E. Final Project Working Screenshots and Results

The project demonstrated success through model performance and visualization assets.

1. Model Performance (Python Results)

The **Random Forest Classifier** achieved the following on the segregated test data:

- **Test Set Accuracy: 1.0000 (100%)**
- **Classification Report Summary:** Precision, Recall, and F1-Score of 1.00 for the class present in the test set.

  Supporting Explanation: This perfect score demonstrates the model's strong ability to learn the patterns within the data and correctly classify the held-out samples. *However, this high score is qualified by the small, less diverse nature of the test set, as noted in the limitations.*

  2. Feature Importance Analysis (Key Finding)

  Supporting Explanation: The Feature Importance chart (often a bar chart in Python) clearly showed that these three chemicals collectively account for over 64% of the model's predictive power, providing a crucial, actionable business recommendation.

  3. Deployment Ready Asset

- **Deliverable:** Screenshot confirming the successful export of random_forest_wine_quality_model.pkl.

  Supporting Explanation: This file is the key deliverable that allows the intelligence to be transferred from the Python environment to the Power BI reporting environment without requiring a manual recalculation, thereby operationalizing the model.

- Project GitHub Link : https://github.com/sure-trust/RITIK-RAUSHAN-g8-ds/tree/main/Final%20capstone%20project

Learning and Reflection

This section provides a critical self-assessment of the internship experience, highlighting key technical and professional growth achieved throughout the project lifecycle.

A. New Learnings Acquired

The project required proficiency across the entire data science stack, leading to significant new learning in several areas:

1. Technical Learning (Technology and Tools)

- **End-to-End Pipeline Integration:** Gained practical mastery in connecting disparate tools—specifically, integrating a clean **SQL View** with a Python modeling environment, and then consuming the serialized **.pkl model** file directly within **Power BI**. This provided essential, real-world experience in operationalizing a model, a key skill for production data science.

- **Advanced Feature Engineering:** Learned how to move beyond basic data cleaning to create a high-impact, domain-specific feature (RiskScore). This process required deep critical thinking about the chemical interactions within the wine data and their effect on the target variable.

- **Deployment and Serialization:** Mastered the use of the pickle library for model serialization and developed the framework for an **Automation Script**. This was crucial for understanding MLOps principles—how to build a model that is sustainable and ready for continuous integration and continuous deployment (CI/CD).

- **Random Forest Interpretability:** Deepened understanding of the **Random Forest Classifier** by leveraging its intrinsic **Feature Importance** output, which is key for translating complex model behavior into clear, actionable business insights.

2. Professional Learning (Management and Soft Skills)

- **Project Scoping and Time Management:** Learned to define clear **Scope and Limitations** upfront, managing expectations regarding the **100% accuracy** result and focusing efforts on the most critical deliverables, particularly the deployment assets.

- **Stakeholder Communication:** Developed skills in translating complex technical outputs (like the Classification Report) into concise, high-impact business language, as demonstrated in the Executive Summary. Learned how to communicate the strategic value of insights (e.g., *why* focusing on Volatile Acidity matters).

- **Critical Evaluation:** Gained experience in critically evaluating model performance, recognizing that a perfect score on a small test set is a signal for strong learning but also a warning sign for potential overfitting, necessitating caution and future work (Hyperparameter Optimization).

B. Overall Experience and Reflection

The internship project provided an invaluable opportunity to execute a full data science life cycle.

- **Overall Experience:** The experience was highly rewarding, transitioning knowledge from theoretical concepts to a tangible, deployable solution. The most fulfilling aspect was seeing the data flow seamlessly from the initial SQL preparation through to the final, interactive **Power BI Dashboard** .
- **Impact Realization:** The biggest takeaway was understanding that the ultimate value of a data science project is not in the model's accuracy, but in its **actionability**. The engineered RiskScore and the clear Feature Importance insights proved to be the most valuable deliverables for stakeholders, emphasizing the importance of business context over purely technical metrics.
- **Growth:** The project served as a definitive demonstration of proficiency across the modern data science toolkit, solidifying confidence in tackling complex, real-world analytical problems in a professional capacity.

Conclusion and Future Scope

Conclusion: Recap Objectives and Achievements

This internship project successfully delivered the **Wine Quality Intelligence Suite**, achieving its core objectives through a structured data science pipeline:

Objective	**Achievement**	**Impact**
Prediction	Trained a highly performant **Random Forest Classifier** model.	Provides objective, instantaneous quality scoring, minimizing reliance on subjective sensory panels.
Innovation	Successfully engineered the **RiskScore** metric.	Creates a simple, actionable operational metric for risk management and batch intervention.
Insight	Identified **Volatile Acidity** and **Chlorides** as the dominant factors.	Offers clear, prioritized guidance to winemakers for quality improvement efforts.
Deployment	Created the **.pkl model file** and the **Automation Script** template.	Ensures the solution is scalable, maintainable, and ready for integration into the Power BI business intelligence environment.

In summary, the project moved the quality assessment process from a historical, costly evaluation to a real-time, objective intelligence stream, providing a substantial operational and strategic advantage.

B. Future Scope of this Project

To evolve this solution from a successful prototype into a robust, enterprise-grade system, the following actions are recommended for future development:

1. Predictive Modeling Enhancement

- **Hyperparameter Optimization:** Conduct rigorous tuning (e.g., using Grid Search or Bayesian Optimization) on the Random Forest parameters to find the optimal balance and improve generalization, especially across under-represented quality classes.

- **Comparative Modeling:** Evaluate high-performance alternatives, such as **XGBoost** or **LightGBM**, which are highly effective for structured data classification, to benchmark the performance against the current Random Forest.
- **Addressing Imbalance:** Implement advanced techniques like **SMOTE** (Synthetic Minority Over-sampling Technique) or cost-sensitive learning to improve the model's ability to accurately predict the rare, high-value (Quality 7/8) and low-value (Quality 3/4) wines.

2. Deployment and Integration Expansion

- **Real-Time Scoring API:** Deploy the serialized model on a cloud platform (e.g., Azure Functions or AWS Lambda) as a REST API. This would allow new data from lab equipment to be sent to the API for an instant prediction, enabling true *real-time* quality monitoring.
- **Interactive Web Application:** Develop a user-friendly front-end application (using frameworks like Streamlit or Flask) that allows non-technical staff to manually input chemical readings and receive an instant quality prediction and risk level, bypassing the need for the Power BI dashboard for quick checks.

3. Data and Feature Expansion

- **External Validation:** Incorporate data from other wine regions or vintages to test the model’s **generalizability** and potentially retrain it on a more diverse, global dataset.
- **Time-Series Analysis:** Introduce time-based features (e.g., age of the wine, month of production) and explore time-series forecasting models to predict *future* quality trends rather than just current quality scores.