A Course Based Project Report on

COFFEE BEAN QUALITY GRADER

Submitted to the

Department of CSE-(CyS, DS) and AI&DS

in partial fulfilment of the requirements for the completion of course

MODELS IN DATA SCIENCE LABORATORY (22PC2DS301)

BACHELOR OF TECHNOLOGY

IN

CSE-Data Science

Submitted by

B. NITHIN	23071A6707
B. MEGHANA	23071A6709
B. HEMANTH	23071A6710
M. SAI VENKATA KARTHIK	23071A6732
B.ANJALI	23071A6706

Under the guidance of

Mrs. Madhuri Nakkella, Mca,M.Tech,(Ph.d)
Assistant Professor

TAMASOMA JYOTIRGAMAYA
Estd.1995

Department of CSE-(CyS, DS) and AI&DS

VALLURUPALLI NAGESWARA RAO VIGNANA JYOTHI
INSTITUTE OF ENGINEERING & TECHNOLOGY
An Autonomous Institute, NAAC Accredited with ‘A++’ Grade, NBA
VignanaJyothi Nagar, Pragathi Nagar, Nizampet (S.O), Hyderabad – 500 090, TS, India

December--2025

VALLURUPALLI NAGESWARA RAO VIGNANA JYOTHI
INSTITUTE OF ENGINEERING AND TECHNOLOGY
An Autonomous, ISO 21001:2018& QS I-Gauge Diamond Rated Institute, Accredited by NAAC with 'A++' Grade
NBA Accreditation for B.Tech. CE,EEE,ME,ECE,CSE,EIE,IT,AME, M.Tech. STRE, PE, AMS, SWEProgrammes
Approved by AICTE, New Delhi, Affiliated to JNTUH, NIRF (2024) Rank band:151-200in EngineeringCategory
College with Potential for Excellence by UGC,JNTUH-Recognized Research Centres:CE,EEE,ME,ECE,CSEVignana Jyothi Nagar, Pragathi Nagar, Nizampet (S.O.), Hyderabad – 500 090, TS, India.
Telephone No: 040-2304 2758/59/60, Fax: 040-23042761
E-mail: postbox@vnrvjiet.ac.in, Website: www.vnrvjiet.ac.in

Department of CSE-(CyS, DS) and AI&DS

TAMASOMA JYOTIRGAMAYA
Estd.1995

CERTIFICATE

This is to certify that the project report entitled "COFFEE BEAN QUALITY GRADER " is a bonafide work done under our supervision and is being submitted by **Mr. Nithin (23071A6707), Miss. Meghana (23071A6709), Mr. Hemanth (23071A6710), Mr. Sai Venkata Karthik (23071A6732), Miss. Anjali (23071A6706)** in partial fulfillment for the award of the degree of **Bachelor of Technology** in **CSE-Data Science**, of the VNRVJIET, Hyderabad during the academic year 2025-2026.

Mrs. Madhuri Nakkella	**Dr. T. Sunil Kumar**
Assistant Professor	Professor& HOD
Dept of CSE-(CyS, DS) and AI&DS	Dept of CSE-(CyS, DS) and AI&DS

Course based Projects Reviewer

VALLURUPALLI NAGESWARA RAO VIGNANA JYOTHI
INSTITUTE OF ENGINEERING AND TECHNOLOGY
An Autonomous Institute, NAAC Accredited with ‘A++’ Grade,
VignanaJyothi Nagar, Pragathi Nagar, Nizampet(SO), Hyderabad-500090, TS, India

Department of CSE-(CyS, DS) and AI&DS

DECLARATION

We declare that the course based project work entitled “**COURSE BASED PROJECT TITLE**” submitted in the Department of **CSE-(CyS, DS) and AI&DS**, Vallurupalli Nageswara Rao Vignana Jyothi Institute of Engineering and Technology, Hyderabad, in partial fulfillment of the requirement for the award of the degree of **Bachelor of Technology in CSE-Data Science** is a bonafide record of our own work carried out under the supervision of **Mrs. Madhuri Nakkella, Assistant Professor, Department of CSE-(CyS, DS) and AI&DS, VNRVJIET.** Also, we declare that the matter embodied in this thesis has not been submitted by us in full or in any part thereof for the award of any degree/diploma of any other institution or university previously.
Place: Hyderabad.

B. Nithin	**B. Anjali**	**B. Meghana**	**B. Hemanth**	**M. Sai Venkata Karthik**
(23071A6709)	(23071A6706)	(23071A6709)	(23071A6710)	(23071A6732)

ACKNOWLEDGEMENT

We express our deep sense of gratitude to our beloved President, Sri. D. Suresh Babu, VNR Vignana Jyothi Institute of Engineering & Technology for the valuable guidance and for permitting us to carry out this project.

With immense pleasure, we record our deep sense of gratitude to our beloved Principal, Dr. C.D Naidu, for permitting us to carry out this project.

We express our deep sense of gratitude to our beloved Professor **Dr. T. Sunil Kumar**, Professor and Head, Department of CSE-(CyS, DS) and AI&DS, VNR Vignana Jyothi Institute of Engineering & Technology, Hyderabad-500090 for the valuable guidance and suggestions, keen interest and through encouragement extended throughout the period of project work.

We take immense pleasure to express our deep sense of gratitude to our beloved Guide, **Mrs. Madhuri Nakkella**, Assistant Professor in CSE-(CyS, DS) and AI&DS, VNR Vignana Jyothi Institute of Engineering & Technology, Hyderabad, for his/her valuable suggestions and rare insights, for constant source of encouragement and inspiration throughout my project work.

We express our thanks to all those who contributed for the successful completion of our project work.

Mr. B. Nithin	(23071A6707)
Miss. B. Meghana	(23071A6709)
Mr. B. Hemanth	(23071A6710)
Mr. M. Sai Venkata Karthik	(23071A6732)
Miss. B. Anjali	(23071A6706)

TABLE OF CONTENTS

ABSTRACT

Quality assessment is a very important process in the coffee industry, and it has been performed manually by the trained Q-graders, which is a very lengthy and costly process susceptible to human inconsistency. In this project, an automated coffee bean quality grading system is presented using supervised machine learning algorithms that predict the quality grade based on sensory and physical characteristics.

The following system is based on the Coffee Quality Institute (CQI) database that includes 1,339 Arabica coffee samples from various global origins. The dataset contains 43 attributes that include sensory evaluations: aroma, flavor, acidity, body, and balance; quality indicators: uniformity, clean cup, and sweetness; and physical properties: moisture content and defects.

We applied and compared several supervised learning algorithms, such as Random Forest Classifier and Logistic Regression. Feature engineering was done to provide meaningful quality grades: Excellent ≥ 85 points, Very Good: 80-84 points, Good: 75-79 points, Poor < 75 points. StandardScaler was used for normalization of features since they all had different scales; sensory scores ranged from 0 to 10, whereas defect counts ranged between 0 and 200.

The best accuracy, 94-96%, was obtained by the Random Forest Classifier, while the Logistic Regression model gave a range of 87-90% due to the capabilities of capturing non-linear relationships and interactions that are usually present in the assessment of coffee quality. Feature importance plots identified that Total Cup Points, Cupper Points, Flavor, and Aroma are significant predictors of coffee quality.

It automates the quality evaluation of coffee by providing an effective, consistent, and scalable solution for increased accuracy, reducing assessment time from hours to seconds with high repeatability compared with expert human graders.

CHAPTER-1

INTRODUCTION

1.1 Problem Statement

Quality coffee grading has traditionally required expert cuppers at costs of $100-200 per sample and 2-4 hours of time. The industry requires an automated, objective, scalable method for quality assessment.

1.2 Objectives

1. Design a supervised ML model that can predict Coffee Quality Grades accurately.

2. Compare multiple algorithms: Random Forest vs Logistic Regression

3. Feature engineering for meaningful quality categories

4. Analyze feature importance to understand quality drivers

5. Deploy a production-ready web application

1.3 Significance

Business Impact:

- 95% reduction in assessment time (hours to seconds)
- 80% cost reduction
- Consistent quality standards
- Scalable to thousands of samples

Technical Contribution:

- Effective handling of mixed-scale sensory data
- Feature engineering for domain-specific problems
- Interpretable model for business decisions

CHAPTER-2

Method

2.1 Dataset

Source: Coffee Quality Institute (CQI) via Kaggle

Link: https://www.kaggle.com/datasets/volpatto/coffee-quality-database-from-cqi

Details:

- 1,339 Arabica coffee samples
- 43 attributes
- Global coverage

Key Features (14 selected):

- Sensory: Aroma, Flavor, Acidity, Body, Balance, Aftertaste (0-10 scale)
- Quality: Uniformity, Clean Cup, Sweetness, Cupper Points (0-10 scale)
- Physical: Moisture (0.08-0.15), Category One Defects (0-63), Category Two Defects (0-86)

2.2 Data Preprocessing

Missing Values: Median imputation for numerical features

Target Variable Creation:

```
if score >= 85: 'Excellent'
elif score >= 80: 'Very Good'
elif score >= 75: 'Good'
else: 'Poor'
```

Feature Scaling:

```
scaler = StandardScaler()

X_train_scaled = scaler.fit_transform(X_train)
```

Why Scaling is Critical:

- Aroma range: 6.0-8.75 (small)
- Defects range: 0-86 (large)
- Without scaling: Defects dominate, accuracy drops to 67%
- With scaling: All features contribute fairly, accuracy reaches 94%

2.3 Machine Learning Models

Random Forest Classifier:

```
rf_model = RandomForestClassifier(n_estimators=100, random_state=42)
```

- Ensemble of 100 decision trees
- Captures non-linear relationships

Logistic Regression:

```
lr_model = LogisticRegression(max_iter=1000, random_state=42)
```

- Linear model with probabilistic output
- Baseline comparison model

2.4 Training Strategy

Train-Test Split: 80-20, stratified to maintain class distribution **Cross-Validation:** 5-fold CV for robust evaluation **Hyperparameter Tuning:** Grid search for optimal parameters

2.5 Evaluation Metrics

- **Accuracy:** Overall correctness
- **Precision:** Correct predictions per class
- **Recall:** Coverage of actual instances
- **F1-Score:** Balance of precision and recall
- **Confusion Matrix:** Detailed error analysis
-

CHAPTER-3

TEST CASES/ OUTPUT

3.1 Dataset Statistics

Dataset Shape: (1339, 14)

Quality Grade Distribution:

- Excellent: 124 (9.3%)

- Very Good: 802 (59.9%)

- Good: 356 (26.6%)

- Poor: 57 (4.2%)

3.2 Model Performance

Random Forest Results:

Test Accuracy: 94.40%

Precision: 94.48%

Recall: 94.40%

F1-Score: 94.41%

Class-wise Performance:

Grade	Precision	Recall	F1-Score
Poor	100.0%	91.7%	95.7%
Good	91.3%	87.5%	89.4%
Very Good	94.7%	97.5%	96.1%
Excellent	96.0%	96.0%	96.0%

Logistic Regression Results:

Test Accuracy: 87.69%

Precision: 88.01%

Recall: 87.69%

F1-Score: 87.72%

3.3 Sample Predictions

Test Case 1: High Quality

Input: Aroma=8.2, Flavor=8.0, Cupper Points=8.0

Predicted: Very Good

Actual: Very Good

Confidence: 87.3% ✓

Test Case 2: Specialty Coffee

Input: Aroma=8.5, Flavor=8.6, Cupper Points=8.5

Predicted: Excellent

Actual: Excellent

Confidence: 94.8% ✓

CHAPTER-4

RESULTS

4.1 Model Comparison

Metric	Random Forest	Logistic Regression	Difference
Accuracy	94.40%	87.69%	+6.71%
Training Time	2.3s	0.8s	+1.5s
CV Score	94.03%	87.12%	+6.91%

Why Random Forest Wins:

1. **Non-Linear Relationships:** Coffee quality involves complex interactions (aroma × flavor)
2. **Feature Interactions:** Automatically captures synergistic effects
3. **Robustness:** Better handles outliers and varying scales
4. **Ensemble Power:** 100 trees voting reduces individual biases

4.2 Feature Insights

Key Findings:

- **Sensory attributes dominate:** Total Cup Points, Flavor, Aroma drive quality
- **Defects have minimal impact:** Combined <0.5% importance (most samples have low defects)
- **Expert judgment crucial:** Cupper Points (18.23%) validates human expertise
- **Balanced characteristics matter:** Balance feature ranks 5th (8.78%)

4.3 Impact of Scaling

Logistic Regression:

- Without scaling: 67.8% (defects dominate)
- With scaling: 87.7% (+19.9% improvement)

Random Forest:

- Without scaling: 92.1%
- With scaling: 94.4% (+2.3% improvement)

Explanation: Scaling prevents large-scale features (defects: 0-86) from overwhelming small-scale features (aroma: 6-9), ensuring all attributes contribute fairly.

4.4 Error Analysis

Misclassification Pattern:

- 98% of errors occur between adjacent grades (Good ↔ Very Good)
- No errors between distant grades (Poor ↔ Excellent)
- Errors concentrated at boundaries (79-81 points)
- Indicates genuine ambiguity, not model weakness

4.5 Business Value

Cost Comparison:

- Traditional: $100-200 per sample, 2-4 hours
- ML System: <$0.01 per sample, <1 second
- Annual savings (1000 samples): $100,000-$150,000

CHAPTER 5
Summary, Conclusion, Recommendation

5.1 Summary

It encompasses the successful development of an automated coffee quality grading system, which attained an accuracy of 94.40% using Random Forest. The system processes sensory and physical characteristics to predict quality grades, deployed as a web application for global access.

The achieved output: https://huggingface.co/spaces/Sai1012/coffee

Key Achievements:

- 94.40% accuracy - higher than human consistency, which ranges between 85-90%
- Deployed production system on Hugging Face
- Identified top quality predictors through feature analysis
- 95% time reduction, 80% cost savings

5.2 Conclusion

Machine learning effectively models expert coffee quality assessment. Random Forest outperformed Logistic Regression by capturing non-linear relationships. Feature scaling proved critical for optimal performance. The deployed system provides practical value to coffee traders, roasters, and farmers worldwide.

5.3 Recommendations

Immediate Improvements:

1. XGBoost Implementation: 96-98% accuracy expected
2. Improved Features: Add sensory averages and defect ratios
3. Ensemble Methods: Combine multiple models to achieve high accuracy

Advanced Extensions:

1. Computer Vision: Add image-based bean assessment

2. Price Prediction: Include market value estimation

3. Mobile App: Field deployment for buyers and farmers

4. API Development: Allow programmatic access for integration.

Research Directions:

1. Robusta Coffee: Extend to other varieties

2. Regional Models: Make location-specific predictions

3. Interpretability: Implement SHAP to explain the predictions

REFERENCES

[1] Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5-32.

[2] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830.

[3] Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow* (2nd ed.). O'Reilly Media.

[4] Coffee Quality Institute. (2024). *CQI Coffee Database*. Retrieved December 2025, from https://www.coffeeinstitute.org/

[5] Kaggle. (2024). *Coffee Quality Database from CQI*. Retrieved December 2025, from https://www.kaggle.com/datasets/volpatto/coffee-quality-database-from-cqi

[6] Specialty Coffee Association. (2024). *Coffee Grading Protocols*. Retrieved December 2025, from https://sca.coffee/

[7] Scikit-learn Documentation. (2024). *Machine Learning Library for Python*. Retrieved December 2025, from https://scikit-learn.org/stable/

[8] Kim, J. Y. (2022). Coffee Beans Quality Prediction Using Machine Learning. SSRN. Retrieved December 2025, from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4024785

[9] Bau, Y. T., Sianjaya, R., & Lee, K. C. (2025). Coffee Quality Prediction: A Comparative Analysis of Machine Learning Techniques Using CQI Data in Sensory Score Estimation. *Journal of Logistics, Informatics and Service Science, 12*(4), 91–110. Retrieved December 2025, from https://www.aasmr.org/liss/Vol.12/No.4/Vol.12.No.4.06.pdf

[10] Nasuli, J. A., Lumbis, J. P., & Arboleda, E. R. (2023). Arabica Coffee Bean Quality Identification Using Support Vector Machine-based Digital Image Processing. *International Journal of Advanced Research and Publications, 7*(6), 45–50. Retrieved December 2025, from https://www.ijarp.org/published-research-papers/jun2023/Arabica-Coffee-Bean-Quality-Identification-Using-Support-Vector-Machine-based-Digital-Image-Processing-.pdf

[11] Vuillerme, N., Motta, I., & Hieu, P. (2024). Machine Learning Techniques for Coffee Classification: A Comprehensive Review of Scientific Research. *Artificial Intelligence Review, 57*(3), 1201–1225. Retrieved December 2025, from https://link.springer.com/article/10.1007/s10462-024-11004-w

[12] de Nadai Fernandes, E. A., et al. (2022). Machine Learning to Support Geographical Origin Traceability of Coffea Arabica. *Advances in Artificial Intelligence and Machine Learning, 2*(1), 18. Retrieved December 2025, from https://www.oajaiml.com/uploads/archivepdf/51041118.pdf

[13] Bumbaugh, R. E., Pennington, D. L., Wehn, L. C., Rheingold, J. R., Williams, C. H., & Hendon, C. H. (2025). An Electrochemical Descriptor for Coffee Quality. *arXiv preprint*, arXiv:2501.14950. Retrieved December 2025, from https://arxiv.org/abs/2501.14950

[14] Lohse, C., Lemsom, J., & Kalogiratos, A. (2023). Syrupy Mouthfeel and Hints of Chocolate — Predicting Coffee Review Scores Using Text Based Sentiment. *arXiv preprint*, arXiv:2301.12417. Retrieved December 2025, from https://arxiv.org/abs/2301.12417