Predictive Dunking: Machine Learning Insights into Biscuits Softening Dynamics in Tea

The objective of this project is to conduct an in-depth analysis of provided data related to structural properties of different types of biscuits. The project entails using machine learning algorithm and model optimization techniques to explore the data and derive meaningful insights. It aims to compare different approaches to understanding the data and to draw conclusions based on analytics.

ABSTRACT

This study delves into the structural properties of biscuits and their dynamic interactions with tea during the dunking process, employing advanced machine learning algorithm and meticulous model optimisation techniques. Through the analysis of three comprehensive datasets, each shedding light on distinct facets of biscuits-tea interaction, the research showcases the prowess of machine learning in accurately discerning biscuit type based on their inherent structural characteristics. Moreover, the investigation extends to evaluating the predictive accuracy of Washburn equation in delineating biscuit behaviour, juxtaposed with the performance of machine learning regressors. In addition to classification endeavours, the study endeavours to distinguish between various biscuit types while comparing their respective pore radii. Leveraging python libraries, including Random Forest and Support Vector Machine (SVM) algorithms, the research meticulously analyses experimental parameters to unravel intricate patterns. Furthermore, the application of probability optimisation technique serves to argument the precision of the model. Th report meticulously outlines the methodological, approach, engages in comprehensive discussion of the findings, and offers profound insights into future avenues for data driven exploration in the realm of biscuits science and manufacturing.

INTRODUCTION

The dunking biscuits into tea represents a quintessential ritual in many cultures, embodying a delicate interplay between culinary tradition and sensory experience. Yet, beneath this seemingly simple act lies a complex scientific phenomenon, wherein the structural properties of biscuits intricately interact with the physical properties of the surrounding liquid. Understanding this phenomenon not only holds relevance for gastronomic enthusiasts but also bears significant implications for the food industry, particularly in the realm of biscuit manufacturing.

The structural composition of biscuits, characterised by interlocking gluten fibres and porous microstructures, gives rise to a myriad of dynamics during dunking process. As the biscuit contacts the tea, capillary action facilitates the ingress of liquid into its porous matrix, leading to observable changes in properties such as texture, flavour, and structural integrity. However, the extent to which these interactions vary across different types of biscuits remains a subject of intrigue and scientific inquiry. Before delving into the intricacies of our analytical endeavours, it is imperative to acquaint ourselves with the datasets at our disposal. We are presented with three distinct datasets, each encapsulating unique facets of the biscuit-tea interaction phenomenon: the dunking dataset, microscopy-data, and the time-resolved (TR) data, further divided into tr-1, tr-2, and tr-3.

The dunking dataset serves as the cornerstone of our analysis, encompassing a comprehensive array of experimental parameters essential for understanding biscuit behaviour during the dunking process. These parameters include gamma, representing the tea surface tension in Nm−1; phi, denoting the contact angle between the biscuit and the tea surface in radians; eta, indicative of the tea dynamic viscosity in Pas; L, signifying the distance up the biscuit that the tea is visible in meters; t, depicting the time elapsed after the initial dunking when the measurement was recorded in seconds; and finally, the biscuit type, encompassing variants such as Rich Tea, Hobnob, or Digestive. Leveraging this dataset, our objective is to harness the power of machine learning algorithms, specifically Random Forest (RF) and Support Vector Machine (SVM), to accurately classify the various types of biscuits based on their structural properties.

Transitioning to the microscopy-data, we encounter a subset derived from the dunking dataset, where the focus shifts to a microscopic examination of biscuit pore radii. This dataset, devoid of biscuit type information but enriched with pore radius data, offers invaluable insights into the structural nuances distinguishing between biscuit varieties. Through meticulous analysis and visualization techniques, we endeavour to elucidate how pore radius differs among the three biscuit types, thus shedding light on an understudied aspect of biscuit morphology.

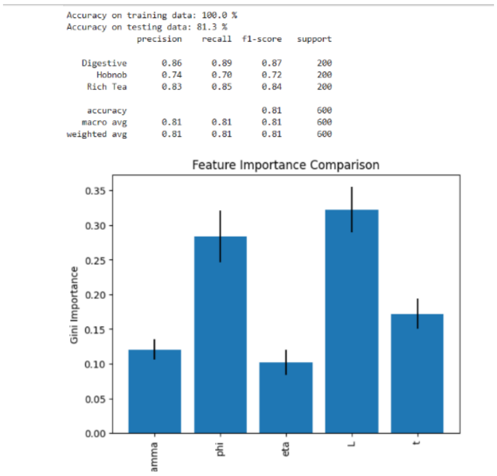
Finally, the time-resolved (TR) datasets, namely tr-1, tr-2, and tr-3, present a nuanced perspective on biscuit-tea interaction dynamics over varying time intervals. These datasets measure the length of tea absorption into biscuits over a time range spanning from 30 seconds to 300 seconds, with gamma, phi, and eta held constant. Notably, while each TR dataset captures measurements for different biscuit types, the specific biscuit utilized for each measurement remains undisclosed. To glean meaningful insights from this data, we leverage optimization techniques, particularly focusing on the uncertainty parameter represented by dL, to discern optimal modelling strategies and refine our understanding of biscuit-tea interaction kinetics.

In summary, the synthesis of these diverse datasets forms the bedrock of our analytical framework, enabling us to unravel the intricate nuances of biscuit-tea interaction and pave the way for advancements in biscuit science and manufacturing.

ANALYSIS AND DISCUSSION

Using ML to Identify Various Biscuit Types

This program trains a Random Forest model to predict the type of biscuit based on factors. It preprocesses the data by standardizing characteristics. Divides it into training and testing datasets. After training the model it assesses its performance, on both sets presenting accuracy results. It then showcases the importance of features through bar graphs indicating how crucial each feature is in predicting the cookie type. The Random Forest model achieved an accuracy of 81.3% on the testing dataset. Precision values for Digestive, Hobnob and Rich Tea cookies were 0.86, 0.74 and 0.83 respectively showing rates of predictions for each category. The recall scores were 0.89, 0.70 and 0.85 respectively demonstrating how well the model identifies instances from each category correctly. F1 scores that balance precision and recall were found to be 0.87, 0.72 and 0.84, across all categories analysed in this study.

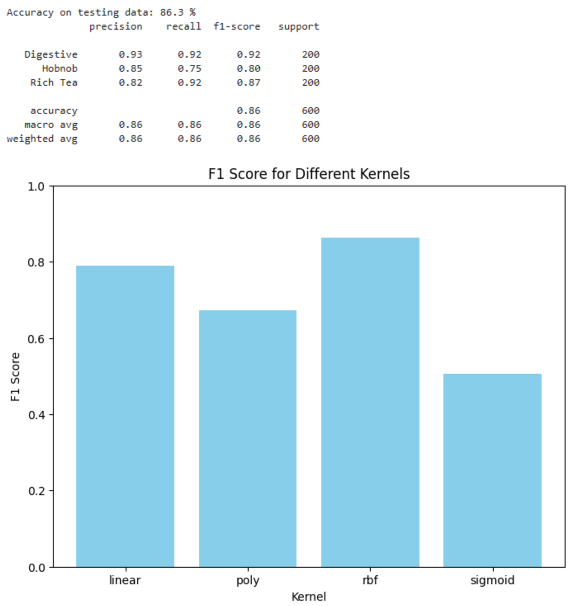


```
Accuracy on training data: 100.0 %
Accuracy on testing data: 81.3 %
              precision    recall  f1-score   support

   Digestive       0.86      0.89      0.87       200
      Hobnob       0.74      0.70      0.72       200
    Rich Tea       0.83      0.85      0.84       200

    accuracy                           0.81       600
   macro avg       0.81      0.81      0.81       600
weighted avg       0.81      0.81      0.81       600
```

Figure 1. Feature Importance Comparison of Random Forest Classifier



```
Accuracy on testing data: 86.3 %
              precision    recall  f1-score   support

   Digestive       0.93      0.92      0.92       200
      Hobnob       0.85      0.75      0.80       200
    Rich Tea       0.82      0.92      0.87       200

    accuracy                           0.86       600
   macro avg       0.86      0.86      0.86       600
weighted avg       0.86      0.86      0.86       600
```

Figure 2. Comparison between different kernel in SVM

Another ML algorithm used was Support Vector Machine, this code evaluates the performance of a Support Vector Classifier (SVC) using different kernel types (linear, polynomial, radial basis function, and sigmoid) on the testing data. It first trains an SVC with a default kernel (RBF) and calculates its accuracy. Then, it iterates through each kernel type, trains a new SVC for each kernel, predicts labels for the test set, calculates the F1 score for each kernel, and finally visualizes the F1 scores for different kernels using a bar plot. The analysis indicates that the RBF kernel achieved the highest accuracy on the testing data (86%) among the tested kernels. It demonstrated consistent performance across precision, recall, and F1-score metrics for each class, indicating good generalization capability. The linear kernel followed with an accuracy of 79%, performing relatively well for the 'Rich Tea' class but less so for others. However, the polynomial and sigmoid kernels showed poorer performance, with lower accuracy and less consistent precision, recall, and F1-scores across classes. Therefore, the RBF kernel is recommended for this classification task due to its superior performance.

In the comparison both classifiers show performance, in predicting the types of biscuits using the settings. The SVM classifier with the RBF kernel performs better than the Random Forest classifier in terms of accuracy on the test data (86% versus 81.3%). For all three categories (Digestive, Hobnob, Rich Tea) the SVM classifier generally demonstrates precision, recall and F1 scores compared to the Random Forest classifier. Therefore, based on these findings it seems that the SVM classifier with the RBF kernel is more effective at predicting biscuit types based on parameters, than the Random Forest Classifier.

Assessing Washburn Equation Accuracy and Comparing with ML regressor

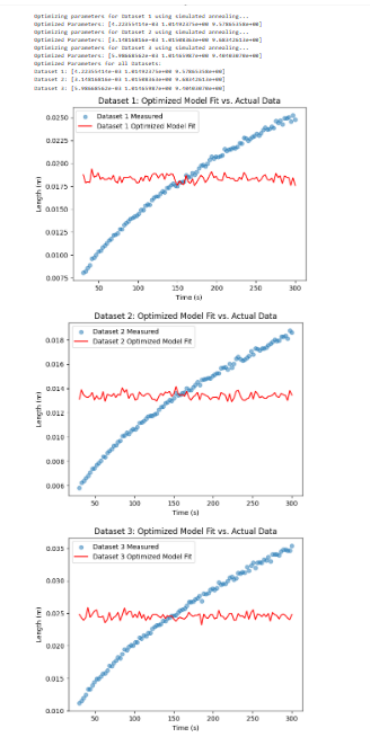
Using the *‘microscopy-data’* the code performs regression analysis on the data using two approaches: a Random Forest Regressor and the Washburn equation. It splits the data into training and testing sets and then trains Random Forest Regressor on the training data and evaluates its performance using mean squared error (MSE) and R-squared (R2) score on the testing data. It also calculates the predicted values using the Washburn equation based on testing data’s input parameters. The result shows that the Washburn equation outperforms the Random Forest Regressor in predicting L values based on experimental parameters, as evidence by its lower MSE and higher R-squared value, it yields more accurate prediction of L using the experimental parameters.

Comparing Pore Radius in Different Biscuit Types

The predictions made by the Random Forest Regressor model earlier is used to predict biscuit type based on pore radius data. It predicts the biscuit types using the trained model and then it classifies the predicted biscuit types into three classes using defined thresholds. Finally, it computes pore radius statistics for each biscuit type, including mean, standard deviation, etc. By the analysis it was observed that type 3 biscuits tend to have larger average pore size compared to type 1 and type 2 biscuits, as indicated by their higher mean pore size.

Model Optimisation

The ‘tr-data’ includes the uncertainty in the length due to which the data is introduced optimisation approach. Using the ‘basinhopping’ algorithm it finds the optimal parameters for the Washburn equation based on the experimental data from multiple ‘tr’ datasets. It defines the negative log-likelihood function that quantifies the difference between predicted length using the Washburn equation and actual length from the datasets. The algorithm iteratively adjusts parameters to minimize this difference, aiming to optimize the model’s fit to the experimental data.



```
Optimizing parameters for Dataset 1 using simulated annealing...
Optimized Parameters: [4.22355415e-03 1.01492375e+00 9.57865358e+00]
Optimizing parameters for Dataset 2 using simulated annealing...
Optimized Parameters: [3.14816816e-03 1.01508363e+00 9.68342613e+00]
Optimizing parameters for Dataset 3 using simulated annealing...
Optimized Parameters: [5.98668562e-03 1.01465987e+00 9.40403870e+00]
Optimized Parameters for all Datasets:
Dataset 1: [4.22355414e-03 1.01492375e+00 9.57865358e+00]
Dataset 2: [3.14816816e-03 1.01508363e+00 9.68342613e+00]
Dataset 3: [5.98668562e-03 1.01465987e+00 9.40403870e+00]
```

Figure 3. Optimised Model fit for tr-1, tr-2, tr-3 data

DISCUSSION

In Summary, the Washburn equation s outperformance of the Random Forest Regressor can be attributed to its physical based foundation, simplicity, and precision in modelling fluid dynamics. The equation's direct relationship between experimental parameters and length predictions eliminates under relevant conditions. However, both model's performance may be affected by factors such as data quality and quantity. Further exploration and optimisation of the dataset could enhance more performance, but time constraints may limit the depth of analysis in this regard.

CONCLUSION

In conclusion, the analysis demonstrates the effectiveness of machine learning techniques in various scenarios, such as identifying biscuit types and predicting pore radius. The SVM classifier with the RBF kernel outperforms the Random Forest Regressor in biscuit type prediction, while the Washburn equation surpasses the Random forest Regressor in predicting length values. However, further exploration of the 'tr ' dataset and optimisation of models could enhance the accuracy of prediction. The constraints limited the depth of analysis, suggesting avenues for future research to delve deeper into the dataset and refine the model for improved performance.