Journal of Advanced Research in Applied Sciences and Engineering Technology

Journal homepage:
https://semarakilmu.com.my/journals/index.php/applied_sciences_eng_tech/index
ISSN: 2462-1943

Statistics and Machine Learning Based Decision-Making for Diet Beverage Choice and Recommendation through Software Application

Arundhuti Chakraborty[1], Santanu Mandal[1,*], Rajkanwar Singh[2]

[1] School of Advanced Sciences, VIT-AP University, Andhra Pradesh, India
[2] School of Computer Science and Engineering, VIT-AP University, Andhra Pradesh, India

ABSTRACT

Positioned within the realm of ubiquitous beverage chains acclaimed for their diverse coffee and tea offerings, this investigation deeply explores the intricate nutritional dynamics and caloric compositions of these widely consumed beverages. At its core, the research aims to streamline decision-making for individuals actively seeking health-conscious beverage alternatives. This comprehensive endeavour unfolds through an integrated methodology that combines hypothesis testing and other statistical techniques with the establishment of a robust decision support system, complemented by powerful machine learning techniques. Distinctively, the paper integrates an intuitive React application (ReactApp) into the robust Fast API framework. Users engage seamlessly with the decision support system, as Fast API efficiently manages data processing, interfaces with machine learning algorithms, and delivers personalized recommendations to the ReactApp front. In essence, this interdisciplinary initiative epitomizes the fusion of nutritional science, statistical analysis, machine learning, and modern web technologies, providing a holistic and pioneering solution for selecting health-conscious beverage alternatives. One of the most noteworthy outcomes of our research lies in the compelling results we've achieved. Our predictive model has demonstrated exceptional performance: We achieved an impressive Accuracy Rate of 83.67%, signifying the high precision of our recommendations in identifying health-conscious beverage alternatives. Simultaneously, the F1-Score, which harmonizes precision and recall, stands at a commendable 82.21%, indicating a well-balanced and effective decision support system. Our model's ability to recall health-conscious choices is also strong, with a Recall rate of 83.67%. Furthermore, we assessed the model's agreement with observed data using Cohen's Kappa, which yielded a substantial score of 78.01%. This indicates not only the model's predictive power but also its ability to capture agreement beyond what might occur by chance alone. These results, seamlessly integrated into our research narrative, underscore the effectiveness and reliability of our approach. They constitute a rare and valuable contribution to the field, promising a novel pathway for informed and health-driven consumer choices, firmly supported by our rigorous methodology and advanced machine learning outcomes.

Keywords:

Statistical analysis; machine learning; decision making; software applications; diet beverages

1. Introduction

* *Corresponding author.*
E-mail address: santanu.mandal@vitap.ac.in

https://doi.org/10.37934/araset.57.2.2840

In the contemporary health-conscious landscape, the consumption of beverages naturally containing caffeine, such as coffee and tea, has become widespread and enduring [1]. However, the regulation of beverages with added caffeine has proven intricate [2]. Amidst this backdrop, individuals increasingly prioritize their well-being, adopting diverse dietary patterns to sustain healthier lifestyles. While attention is often focused on avoiding high-calorie fast foods and sugary drinks, the caloric content of beloved coffee and tea beverages is sometimes overlooked. Prominent beverage chains offer an extensive range of specialized coffee and tea options, varying widely in caloric impact [3]. Despite loyalty to these brands, the abundance of customization choices can overwhelm customers, causing them to unintentionally disregard their overall calorie intake [4].

Examining the contemporary landscape, a rising prevalence of health-conscious individuals underscores the importance of understanding dietary choices [5]. Centres for Disease Control and Prevention data indicates a growing number of adults actively pursuing weight management [6], signalling a heightened awareness of obesity-related health risks and the availability of healthier food alternatives. Amidst this context, numerous popular coffeehouse chains cater to diverse palates with an array of beverages, snacks, and treats [7]. Although customers exhibit brand loyalty, many remain in pursuit of well-informed choices that align with their health objectives [8]. Consequently, a comprehensive examination of the caloric composition of beverages of popular chains is imperative. While the average caloric content hovers around 250 calories, individual beverages can fluctuate significantly. Customization options introduce an added layer of complexity, enabling minor modifications to exert a notable influence on calorie intake [9].

In an effort to simplify the decision-making process for the health-conscious, we have developed a decision support system. Previously, a decision support system using the Weighted Product (WP) method was implemented by Imam *et al.,* [10] for optimal drink selection in Mataram City, considering criteria such as price, composition, type, and size. This research did not use any statistical testing or machine learning approach for the decision making and no information on user interface was discussed. Likewise, another investigation by Faradillah *et al.,* [11] utilized the AHP method to construct an online chosen decision support system for a tea franchise outlet, presenting a potentially feasible solution for potential franchisees. They have not studied the high-accuracy machine learning method for such decision-making. Viejo *et al.,* [12] described the quality and consumer preferences based on machine vision. Whereas our research takes manufacturer data to recommend user choices without using machine vision. In the recent era, extensive research is going on for applications in food industries [13-15]. Now we know that machine learning is a useful technology for decision support systems and assumes greater importance in research and practice [16, 17]. It is also an essential approach to improving dynamic decision-making [18]. Also, hypothesis testing has been instrumental in decision-making processes [19].

Based on the above literature study, in the pursuit of constructing a robust and effective methodology and advancements of machine learning and data analytics, a dual approach was meticulously amalgamated.

i. Initially, the Analysis of variance (ANOVA) methodology was employed as a hypothesis testing mechanism to discern the more influential target variable among the diverse attributes. This strategic selection process served as a crucial step in identifying the pivotal factors contributing to the overall variance.
ii. Subsequently, considering the intricate challenge posed by the presence of multiple labels, the random forest approach was adeptly harnessed. This ensemble technique not

only exhibited remarkable predictive capabilities but also accommodated the complex classification requirements, yielding insightful and accurate results. It is noteworthy that studies integrating such a comprehensive framework are relatively scarce.

The synergistic fusion of ANOVA and random forest techniques represents a novel and innovative approach, capitalizing on the strengths of both methodologies. This amalgamation allows for a more nuanced exploration of the underlying patterns within the data and fosters a deeper understanding of the intricate relationships among variables. As such, this study contributes to the limited body of research that leverages the synergies of these distinct yet complementary techniques to address multifaceted challenges in predictive modelling and decision-making processes. Moreover, this research work adopts a cutting-edge technological approach by seamlessly integrating a React application (ReactApp) with the robust FastAPI framework. The user-friendly ReactApp interface empowers users to effortlessly engage with the decision support system, enabling the effortless input of preferences and providing seamless access to personalized recommendations. Operating as the backend, FastAPI adeptly manages intricate data processing, facilitates fluid communication with advanced machine learning algorithms, and skilfully delivers tailored suggestions to the front end of ReactApp.

This study embarks on a comprehensive exploration, commencing with an inquiry into the interplay of nutritional attributes across diverse beverages. The subsequent application of Principal Component Analysis (PCA) contributes to model optimization, effectively distilling complex data dimensions. Subsequently, the ANOVA technique unveils intrinsic variability within categorical variables, a pivotal guidepost for subsequent modelling choices. Drawing on this insight, our dataset is partitioned into training and testing subsets, seamlessly integrated into a Random Forest model, renowned for its decision-making prowess. This model forms the predictive core of our approach. Additionally, to rigorously assess our model's performance, we deploy ROC curves, meticulously examining the discriminative abilities of each class within the predictions. These curves provide nuanced insights into the model's efficacy across diverse beverage categories, enhancing the comprehensiveness of our evaluation.

Lastly, the apex of our endeavour is epitomized by a user-friendly interface, crystallizing as a sophisticated software application. This collective odyssey underscores the synergy between data analysis, statistical methodologies, and software engineering, culminating in a comprehensive and practical solution for well-informed beverage selection. By understanding this trend and collecting more data on people’s health and preferences, health beverage companies can explore the implications of supply chain risk management practices on the recommendation and adoption of health-conscious diet beverages, bridging the gap between organizational resilience and individual wellness choices [20]. Furthermore, the integration of life cycle assessment, life cycle costing, and multi-criteria decision-making for food waste composting management, as explored by Abu *et al*., exemplifies the broader trend of integrating diverse methodologies to address complex sustainability challenges [21]. As highlighted in the study by Firdaus *et al*., [22] the importance of driver attention is paramount and while phone calls are a major factor, upon further investigation beverage choice can be correlated to driver attention.

2. Methodology

In the dynamic realm of beverage choices offered by various chains, the plethora of options often engenders confusion, especially in the haste of modern life. Herein emerges a user-friendly tool, a compass of sorts, to streamline the complex task of decision-making. This tool empowers customers to effortlessly input their preferences, spanning from desired calorie intake to nuanced

taste inclinations. By zooming in on specific criteria, individuals can promptly unveil beverages that harmonize with their dietary objectives. This expeditious process not only conserves time but also curbs decision fatigue that can often dampen the dining experience.

The true prowess of this tool shines through its ability to spotlight beverages that precisely meet predefined criteria. By illuminating options aligned with specific nutritional and taste preferences, a subtle yet profound influence is wielded, guiding consumers towards judicious beverage selections. This approach not only simplifies the selection process but also cultivates an environment where health-conscious choices seamlessly intertwine with the fabric of everyday life.

Moreover, our research has unearthed a paramount revelation: the pivotal role of beverage preparation in the decision-making equation. This critical factor wields the most profound impact on the nutritional and caloric composition of beverages. Thus, the tool takes into account not only the nutritional attributes of the ingredients but also the method of preparation, enabling customers to make choices that are both gratifying and health-enhancing.

2.1 Feature Reduction

The accuracy of a machine learning algorithm hinges upon the extent to which features are harnessed and raw data is utilized [23]. The comprehensive dataset hails from Starbucks, a renowned and credible beverage chain. This provenance bolsters the authenticity of our study, further emphasizing the robustness and reliability of our findings. Embedded within the dataset [24] are a diverse array of attributes, each contributing distinct facets to the analysis. Encompassing categorical elements like Beverage_category and Beverage prep, alongside a spectrum of quantitative metrics including calories, total fat, trans fat, saturated fat, sodium, total carbohydrates, cholesterol, dietary fiber, sugars, and protein.

The dataset comprises 242 instances, and within it, the presence of outliers and noise necessitates an initial cleaning process. The pre-processing phase encompasses noise reduction strategies, including outlier detection, normalization, and pooling, aiming to rectify unbalanced data due to potential loss of measurements and missing values in patient examinations. Additionally, this stage involves addressing missing values, a crucial step. Among the available 242 instances in the dataset, some entries contain missing values. While the simplest approach involves ignoring these records, this might not be suitable for smaller datasets. Alternatively, employing algorithms to infer missing data or removing records is considered. Nominal features are completed through various modes, while numerical features' gaps are filled using the mean value. This comprehensive pre-processing ensures the dataset is prepared for subsequent analysis with enhanced reliability. To enhance the data's structure and mitigate the high correlation between certain features, a Principal Component Analysis (PCA) approach is employed. This technique involves merging features with strong correlations into singular components, effectively reducing dimensionality while preserving essential information. By optimizing the dataset's representation, this approach establishes a more efficient and effective foundation for subsequent analysis, enhancing the overall quality of insights derived from the data.

As observed in Figure 1, notable correlation (> 0.6) exists between certain columns, specifically 'total fat (g)', 'trans fat (g)', and 'saturated fat (g)'. In response, we implemented Principal Component Analysis (PCA) as a dimensionality reduction technique. Beginning with the extraction of these columns, a distinct dataframe, labeled df2, is fashioned. By applying PCA with a single component, a novel feature was generated to effectively encapsulate the shared variability within these aforementioned columns. Termed 'Total Fat', this outcome embodies a consolidated metric encompassing the fundamental attributes of individual fat content elements. By reintegrating this

synthesized component into the original data frame (df), a unified representation of interrelated fat content data was formed, facilitating subsequent analysis and elucidation.

Utilizing the Principal Component Analysis (PCA) methodology once again, we extended our efforts to harmonize correlated data attributes, specifically 'Sugar', 'Calorie', and 'Carbohydrate'. This is in concurrence with a study that found that the associations of high added sugar intake were linked to low fiber intake, lower fruit and vegetable consumption, and higher wheat consumption, indicating associations of sugar types with dietary carbohydrate and fat quality for non-naturally occurring sugar [21]. Through PCA, these attributes are synthesized into a singular, composite representation, aptly named 'Calories'. This amalgamated component adeptly captures underlying patterns and shared variances intrinsic to the initial attributes, presenting a refined and concise reflection of the information contained within the original columns.

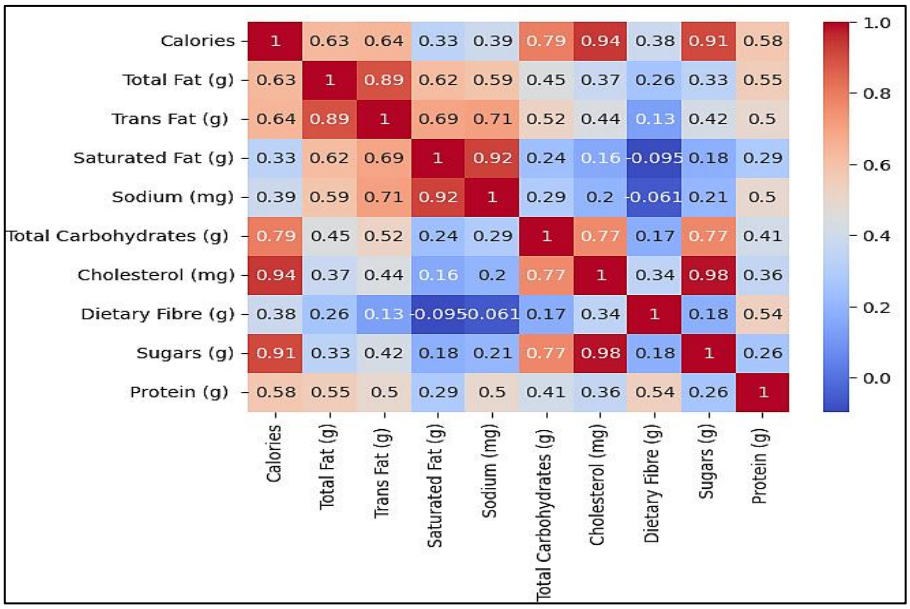

Fig. 1. Correlation heatmap of beverage consumption

2.2 Test for Variability Among Categorical Variables

With two categorical variables, beverage_category and beverage_prep at our disposal, we employed Analysis of Variance (ANOVA) to ascertain the comparative significance between them. This statistical technique enabled us to systematically evaluate the variability among groups and discern whether any statistically significant differences exist. By subjecting the categorical variables to ANOVA, we gained insights into their respective impacts, shedding light on the more influential factor within our analysis.

The analysis conducted using ANOVA based on the data presented in Table 1 has revealed a p-value of less than 0.05 (4.67749937379238e-13), indicating a statistically significant disparity among the mean calorie values associated with distinct beverage preparation methods. This outcome underscores the substantive impact that the selection of a particular beverage preparation method has on the average calorie content of the beverages. The obtained p-value, falling below the predetermined threshold of significance, establishes a robust statistical underpinning for the conclusion that various preparation methods lead to beverages with differing levels of caloric content. Consequently, it can be inferred that the manipulation of beverage preparation techniques stands as a pivotal determinant influencing the caloric composition of beverages within the boundaries of this study.

Table 1
ANOVA results

Index	Sum of squares (sum_sq)	Degrees of freedom (df)	F (F- statistic)	PR (>F)
C(Beverage_category)	1.9776936990740834e-10	2.0	1.5596777299455 39e-14	0.99999987125893 62
C(Beverage_prep)	745062.0275191657	2.0	58.758171313068 594	4.67749937379238 1e-13
C(Beverage_prep):C(Beverage_category)	161255.44792152275	4.0	6.3585814899785 05	0.00036723995283 632796
residual	1489916.8281983421	235.0	NaN	NaN

Upon analyzing the distribution of beverage preparations as depicted in Figure 2, a clear and distinct disparity among the classes becomes apparent. Notably, this disproportionality is particularly evident within the category of milk-based beverages. The initial comprehensive bar plot, encompassing all beverage preparations, effectively portrays the diverse frequencies of the different beverage types. However, a more focused examination uncovers a substantial discrepancy, with milk preparations exhibiting notably higher frequencies in comparison to their non-milk counterparts. This pronounced imbalance accentuates the significance of cautious interpretation and analysis when dealing with data concerning milk-based beverages. Subsequently, a dedicated bar plot exclusively highlighting milk beverage preparations offers a more granular view of this incongruity, facilitating a more precise evaluation of the prevalence of milk-based choices. These visualizations constitute an essential phase in comprehending the distribution patterns within our dataset and hold pivotal importance in our analytical process.

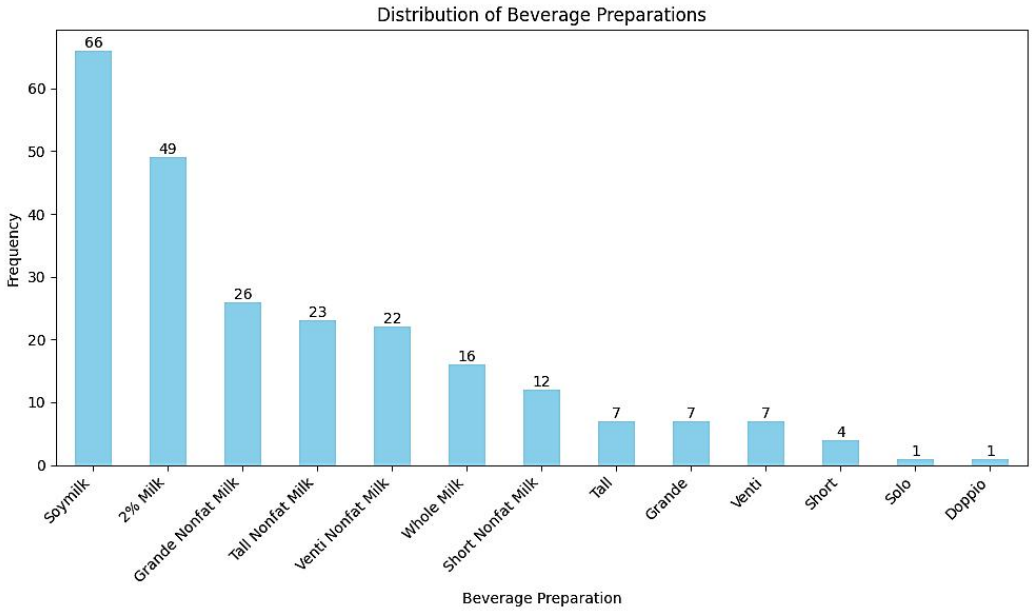

Fig. 2. Distribution of target variable

2.3 Decision-Making through Machine Learning

Given the presence of multiple labels within the target variable, a deliberate strategy was employed involving a diverse ensemble of machine learning techniques, with a specific emphasis on harnessing the capabilities inherent to the Random Forest model. The procedure for selecting the ensemble involved an initial step of feature extraction through Principal Component Analysis (PCA).

Renowned for its robust decision-making process, the Random Forest model effectively demonstrated its competence in capturing intricate relationships within the dataset. By conscientiously choosing and integrating the Random Forest model, our methodology facilitated the establishment of a comprehensive predictive framework. This framework adeptly addressed the complexities presented by the multiple labels in the target variable, culminating in the delivery of precise and insightful analyses. The importance of each feature on a decision tree is then calculated as:

$$ni_j = w_j C_j - w_{left(j)} C_{left(j)} - w_{right(j)} C_{right(j)} \quad (1)$$

where ni_j is the importance of node j, w_j is weighted number of samples reaching node j and C_j= the impurity value of node j.

$$fi_i = \frac{\sum_{j:node\ j\ splits\ on\ feature\ i}^{m} ni_j}{\sum_{k \in all\ nodes}^{1} ni_k} \quad (2)$$

where fi_i is the importance of feature i and ni_j is the importance of node j. These can then be normalized to a value between 0 and 1 by dividing by the sum of all feature importance values:

$$normfi_i = \frac{fi_i}{\sum_{j \in all\ features}^{1} fi_j} \quad (3)$$

In the evaluation of the multiclass classifier's effectiveness, a range of key performance metrics is utilized, offering distinct insights into its competency across various aspects of classification. The F1 score, a harmonic blend of precision and recall, gauges both accuracy and completeness. This is expressed through the formula: F1 = 2 * (precision * recall) / (precision + recall). Accuracy, on the other hand, quantifies the proportion of accurate predictions across all classes, calculated as Accuracy = (TP + TN) / (TP + TN + FP + FN). Recall measures the ratio of accurately classified positive instances to the total actual positives, with the formula: Recall = TP / (TP + FN). True Positive Rate (TPR) mirrors recall in assessing correctly classified positive instances. Conversely, the False Positive Rate (FPR) evaluates the ratio of negative instances erroneously classified as positive, derived as FPR = FP / (FP + TN). Lastly, Kappa captures agreement between raters, factoring in chance agreement through the formula: Kappa = (Observed agreement - Expected agreement) / (1 - Expected agreement). These metrics collectively provide a comprehensive perspective on the classifier's performance.

As depicted in Table 2, in the domain of Random Forest classification, the model's performance emerges as strikingly remarkable. The achieved accuracy, registering at an impressive 83.67%, serves as a testament to its notable predictive proficiency. This measure reflects the ratio of correctly predicted instances to the overall count, showcasing a robust level of predictive accuracy. This competence is further illuminated through the computation of the F1-Score, a metric that harmoniously integrates both precision and recall. With an F1-Score of 0.8221, the model's capacity to classify instances spanning diverse categories precisely and comprehensively of 'beverage preparation' becomes evident. Significantly, the recall metric closely aligns with the observed accuracy, also resting at 83.67%. This alignment reinforces the model's consistent ability to accurately identify positive instances within the context of 'beverage preparation'. The synergy between recall and accuracy emphasizes the model's steadfast performance inappropriately recognizing instances pertaining to the variable of interest. Furthermore, the evaluation is

augmented by the inclusion of Cohen's Kappa, a metric tailored to account for agreement while accommodating class imbalances. The calculated Cohen's Kappa of 0.7801 introduces a nuanced perspective, enhancing the assessment's comprehensiveness. This evaluation is underpinned by the strategic selection of an architecture featuring 75 estimators and the initial partitioning of data into an 80:20 ratio for training and testing. These careful considerations have collectively contributed to the model's exceptional performance, reinforcing its capability to offer accurate and insightful predictions within the intricate landscape of 'beverage preparation' classification.

Table 2
Measures after applying the model

Accuracy	0.8367346938775511
F1-Score	0.8221167254780699
Recall	0.8367346938775511
Cohen's Kappa	0.7801458216489063

The Classification Report, as presented in Table 3, provides a detailed analysis of the model's performance across various classes. Each class corresponds to a specific category of beverage preparation, such as different sizes and milk types. Notably, the '2% Milk' class demonstrates high precision (1.0), indicating accurate predictions, and a substantial recall (0.89), signifying effective identification of true instances. Similarly, the 'Soymilk' class exhibits strong precision (0.87) and recall (0.95). However, some classes, like 'Venti,' lack precision due to inaccurate predictions. Overall, the report offers a comprehensive insight into the model's ability to classify instances across diverse categories, with varying levels of accuracy and recall for each class.

Table 3
Classification report of the target variables

Classification	Precision	Recall	F1-Score	Support
2% milk	1	0.89	0.94	9
Grande nonfat nilk	0.67	0.4	0.5	5
Short nonfat milk	1	1	1	1
Soymilk	0.87	0.95	0.91	21
Tall	0.5	1	0.67	1
Tall nonfat nilk	0.5	0.5	0.5	4
Venti	0	0	0	1
Venti nonfat milk	0.8	1	0.89	4

The utilization of the Receiver Operating Characteristic (ROC) curve here serves to comprehensively evaluate the performance of the binary classification tasks inherent in the multi-class scenario of beverage preparation categories. By plotting the true positive rate against the false positive rate for each class, the ROC curve offers insights into the model's ability to discriminate between different categories. This aids in selecting appropriate classification thresholds, optimizing sensitivity and specificity for individual classes, and comparing model performance. Additionally, the Area Under the ROC Curve (AUC-ROC) provides a consolidated metric summarizing the overall discriminatory power of the model, thereby enabling effective model selection and highlighting strengths and areas of improvement across the diverse beverage preparation classes.

In Figure 3, the displayed ROC curves illuminate the classifier's discriminatory prowess across distinct milk classes. As each ROC curve gravitates towards the upper-left corner, indicating

heightened true positive rates and reduced false positive rates, the classifier's effectiveness becomes increasingly evident.

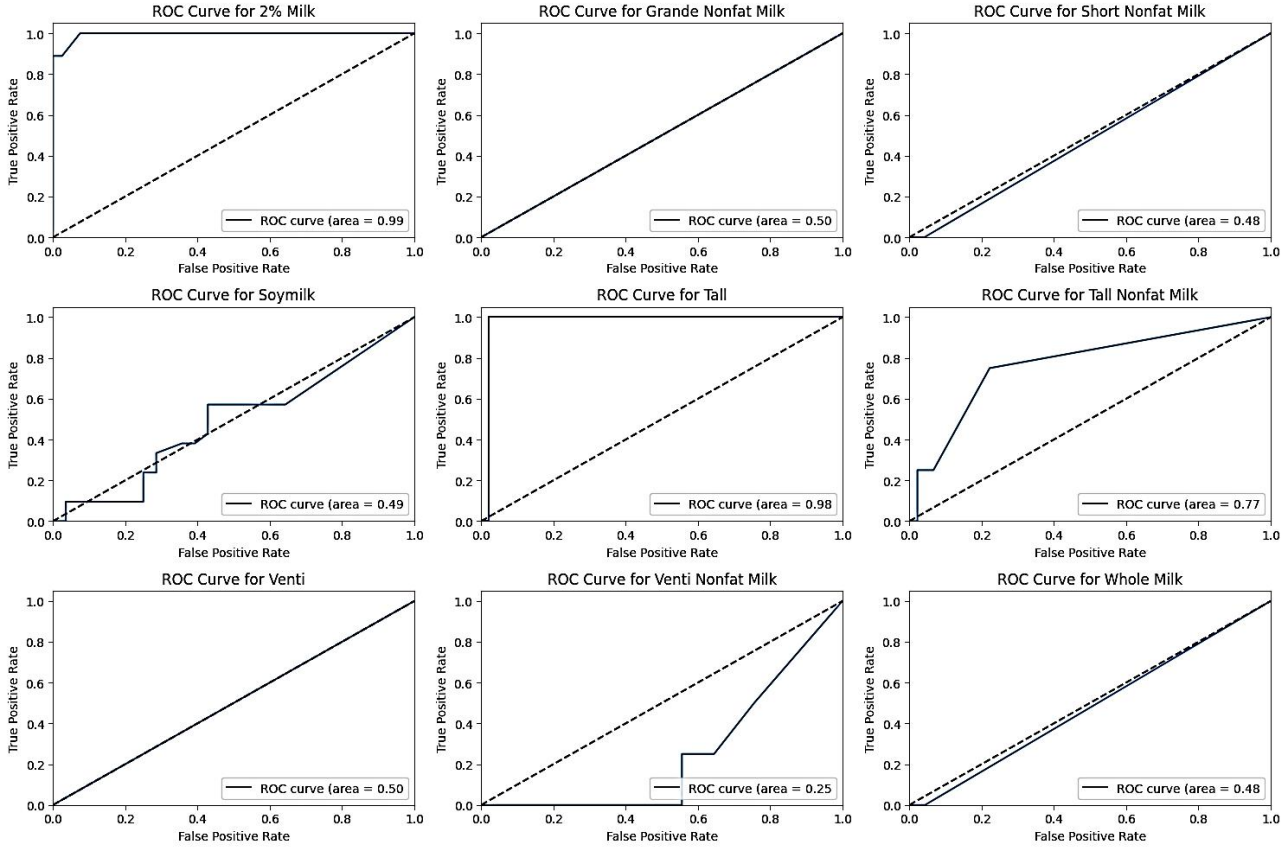

Fig. 3. ROC curves for all the classes of the target variable

Among these curves, the ROC plot corresponding to "2% milk" exhibits the most substantial Area Under the Curve (AUC), underscoring its exceptional ability to differentiate 2% milk from other milk categories. Similarly, the ROC curves for "tall nonfat milk" and "tall milk" display elevated AUC values, signifying their robust discriminatory capacity within their respective class contexts.

Conversely, ROC curves for "grande nonfat milk," "short nonfat milk," "soymilk," "venti nonfat milk," and "whole milk" manifest comparatively lower AUCs, implying a relatively weaker discriminatory performance within their specific class categories.

Taken together, the ROC curve analyses presented in Figure 3 provide a comprehensive overview of the classifier's ability to discriminate across diverse milk types, revealing varying degrees of discriminatory success. notably superior discriminatory performance is observed for "2% milk," "tall nonfat milk," and "tall milk," highlighting their proficiency in class differentiation. this multifaceted evaluation offers valuable insights into the classifier's strengths and areas that could benefit from further refinement.

3. Beverage Recommender: A Software Application

Building upon the comprehensive analysis conducted using the Random Forest classification algorithm, this study introduces a user-friendly application aimed at streamlining the decision-making process for health-conscious individuals in their pursuit of optimal beverage choices. By navigating the intricate landscape of calorie variations and customization options, the application

serves as a robust decision support system. It empowers users to make well-informed and health-conscious decisions aligned with their unique needs and dietary objectives.

At the core of this innovative application is a meticulously trained classification model. This model, characterized by its high accuracy rate and robust performance metrics, excels in accurately categorizing beverages based on their attributes. The result is a more insightful and effective decision-making process for users seeking to align their beverage choices with their health goals.

To bring this application to life, a multi-step implementation process is employed. The trained classification model is serialized and stored for convenient access. Leveraging the FastAPI framework, a dependable backend infrastructure is established, ensuring seamless communication between the user interface and the model. Through a dedicated POST route, users can input their preferences in JSON format, triggering personalized beverage recommendations. The user-friendly interface, developed using React technology, facilitates a smooth interaction between users and the application, ultimately enhancing the overall beverage selection experience.

A visual representation of the application's functionality unfolds in Figure 4. Users engage with the application through a user-friendly interface, entering their preferences and dietary considerations. The application seamlessly processes this input and leverages the power of the Random Forest algorithm to generate tailored beverage recommendations. This process fosters an environment where users can confidently make choices that are both informed and health-driven. In essence, Figure 4 serves as a visual testament to the successful translation of advanced data analysis and machine learning techniques into a practical tool that empowers individuals to make optimal beverage selections. The Random Forest algorithm's prowess in beverage classification merges seamlessly with the application's user-centric design, ultimately enhancing the decision-making experience for health-conscious consumers.

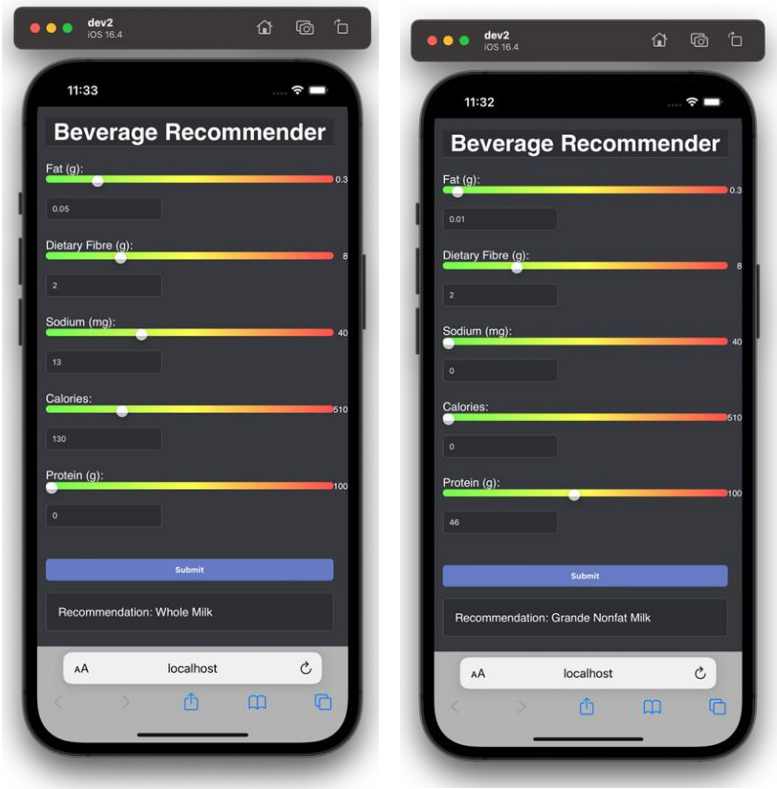

Fig. 4. Beverage recommender as user interface

4. Results

While in previous studies the panel recommended that the consumption of beverages with no or few calories should take precedence over the consumption of beverages with more calories based on generic caloric and nutrient contents and related health benefits and risks [25, 26], our model recommends beverages based on an individual's needs. The flowchart in Figure 5 presents a structured roadmap for crafting an effective calorie and beverage preparation model. Beginning with data import and retrieval, the process advances through pivotal stages. Data splitting into training and test sets, accompanied by scaling, ensures robust model training. Pre-processing tackles data inconsistencies, paving the way for enhanced accuracy.

Principal Component Analysis (PCA) streamlines feature reduction, optimizing data complexity. ANOVA analysis illuminates substantial variations in calorie and beverage attributes, informing feature selection. Exploration of unique preparation types enriches model dynamics.

Central to the model's prowess is the training of a Random Forest algorithm, recognized for its predictive capabilities. The serialized model integrates seamlessly into a FastAPI backend and ReactApp frontend, culminating in a user-friendly application for personalized beverage recommendations. The flowchart exemplifies the fusion of statistical acumen and technological innovation, channelling raw data into a pragmatic solution. This methodology synthesizes intricate data transformations, ultimately empowering health-conscious consumers with informed choices for beverage selection. The flowchart's systematic approach underscores its potential to revolutionize decision-making processes, offering a harmonious blend of analytical rigour and practical utility.

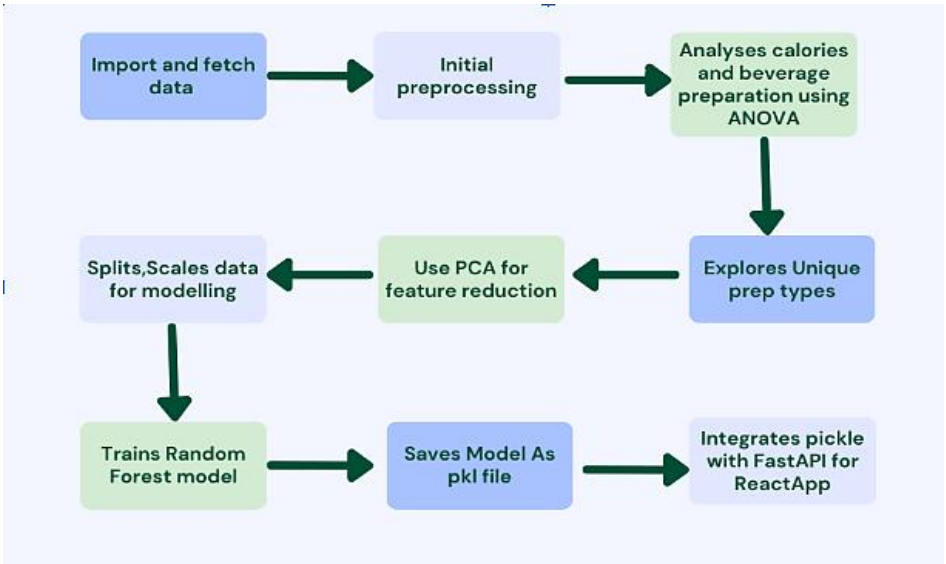

Fig.5. Overview of the process

5. Conclusions

In summary, this study marks a groundbreaking endeavor aimed at simplifying the decision-making process for health-conscious consumers when selecting beverages. By effectively addressing the intricacies of calorie variability and customization preferences, we have developed a

comprehensive decision support system that empowers users to make well-informed and health-conscious choices.

To ensure the quality and integrity of our dataset, we conducted meticulous data preprocessing, including outlier identification, normalization, and handling missing values. Leveraging Principal Component Analysis (PCA), we optimized the data's structure by consolidating correlated features, enhancing its representational efficiency.

Our finely-tuned Random Forest ensemble model, specifically tailored for multiple labels, has demonstrated robust predictive capabilities. With an impressive accuracy rate of 83.67%, an F1-Score of 82.21%, and a kappa of 0.78, the model proficiently categorizes 'beverage preparation' classes. The consistent alignment between recall and accuracy underscores the model's adeptness in identifying positive instances. Notably, this study extends beyond theoretical boundaries, culminating in the creation of a practical and user-friendly React App. This innovative application holds promise for broader contexts, transcending beverages and enabling users across various domains to make health-conscious decisions. As the movement towards health consciousness continues to gain traction, our research contributes to nurturing healthier dietary behaviours and facilitating well-informed choices, thereby enhancing overall well-being and quality of life. It is essential to highlight that our approach is distinctive, as we have successfully integrated both Machine Learning and Statistical ANOVA techniques to build this user interface. Few studies have ventured into this combined approach, further highlighting the uniqueness and potential of our research.

References

[1] McLellan, Tom M., John A. Caldwell, and Harris R. Lieberman. "A review of caffeine's effects on cognitive, physical and occupational performance." *Neuroscience & Biobehavioral Reviews* 71 (2016): 294-312. https://doi.org/10.1016/j.neubiorev.2016.09.001

[2] Flanagan, Robert James, R. A. Braithwaite, S. S. Brown, B. Widdop, F. A. De Wolff, and World Health Organization. *Basic analytical toxicology*. World Health Organization, 1995.

[3] Sovacool, Benjamin K., Morgan Bazilian, Steve Griffiths, Jinsoo Kim, Aoife Foley, and David Rooney. "Decarbonizing the food and beverages industry: A critical and systematic review of developments, sociotechnical systems and policy options." *Renewable and Sustainable Energy Reviews* 143 (2021): 110856. https://doi.org/10.1016/j.rser.2021.110856

[4] Webster, James G. *The marketplace of attention: How audiences take shape in a digital age*. Mit Press, 2014. https://doi.org/10.7551/mitpress/9892.001.0001

[5] SWNS. Over 70 percent of Americans are more health-conscious post-pandemic. New York Post. 2022.

[6] Martin, Crescent B., Kirsten A. Herrick, Neda Sarafrazi, and Cynthia L. Ogden. "Attempts to lose weight among adults in the United States, 2013-2016." *National Center for Health Statistic (NCHS) Data Brief* no. 313 (2018): 1-88.

[7] Goswami, S. "Onerous Journey of a Retail Coffee Outlet Chain (Café La Coffee, India)-(CLC) Imprints of an Expansionist Strategy." *TSM Business Review* 6, no. 1 (2018): 75-87. https://doi.org/10.23837/tbr/2018/v6/n1/174849

[8] Chater, Nick, Steffen Huck, and Roman Inderst. "Consumer decision-making in retail investment services: A behavioural economics perspective." *Report to the European Commission/SANCO* (2010).

[9] Ellison, Brenna, Jayson L. Lusk, and David Davis. "The effect of calorie labels on caloric intake and restaurant revenue: evidence from two full-service restaurants." *Journal of Agricultural and Applied Economics* 46, no. 2 (2014): 173-191. https://doi.org/10.1017/S1074070800000729

[10] Muhammad Imam, Diana, and Rizkillah Muhammad. "Decision Support System to Choose Drinks with The Weight Product (WP) Method." *Jurnal Teknik Informatika CIT Medicom* (2022).

[11] Yanty, Faradillah. "A Decision Support System Of Tea Beverages Outlet Franchise Selection In Indonesia." In *proceedings intl conf information system business competitiveness*. 2012.

[12] Gonzalez Viejo, Claudia, Damir D. Torrico, Frank R. Dunshea, and Sigfredo Fuentes. "Emerging technologies based on artificial intelligence to assess the quality and consumer preference of beverages." *Beverages* 5, no. 4 (2019): 62. https://doi.org/10.3390/beverages5040062

[13] Kumar, Indrajeet, Jyoti Rawat, Noor Mohd, and Shahnawaz Husain. "Opportunities of artificial intelligence and machine learning in the food industry." *Journal of Food Quality* 2021, no. 1 (2021): 4535567. https://doi.org/10.1155/2021/4535567

[14] Sharma, Saurabh, Vijay Kumar Gahlawat, Kumar Rahul, Rahul S. Mor, and Mohit Malik. "Sustainable innovations in the food industry through artificial intelligence and big data analytics." *Logistics* 5, no. 4 (2021): 66. https://doi.org/10.3390/logistics5040066

[15] Ma, Liye, and Baohong Sun. "Machine learning and AI in marketing–Connecting computing power to human insights." *International Journal of Research in Marketing* 37, no. 3 (2020): 481-504. https://doi.org/10.1016/j.ijresmar.2020.04.005

[16] Merkert, Johannes, Marcus Mueller, and Marvin Hubl. "A survey of the application of machine learning in decision support systems." In *23th European Conference on Information System (ECIS)* p. 1-15. 2015. https://doi.org/10.18151/7217429

[17] Bohr, Adam, and Kaveh Memarzadeh. "The rise of artificial intelligence in healthcare applications." In *Artificial Intelligence in healthcare*, pp. 25-60. Academic Press, 2020. https://doi.org/10.1016/B978-0-12-818438-7.00002-2

[18] Meyer, Georg, Gediminas Adomavicius, Paul E. Johnson, Mohamed Elidrisi, William A. Rush, JoAnn M. Sperl-Hillen, and Patrick J. O'Connor. "A machine learning approach to improving dynamic decision making." *Information Systems Research* 25, no. 2 (2014): 239-263. https://doi.org/10.1287/isre.2014.0513

[19] Conteh, Nabie. "The hypothesis testing of decision making styles in the decision making process." *Journal of Technology Research* 1 (2009): 1-17.

[20] Chin, Thoo Ai, and Liu Min. "The effect of supply chain risk management practices on resilience and performance: A systematic literature review." *Journal of Advanced Research in Technology and Innovation Management* 1, no. 1 (2021): 41-53.

[21] Abu, R., Muhammad Arif Ab Aziz, and Zainura Zainon Noor. "Integrated Life Cycle Assessment, Life Cycle Costing and Multi Criteria Decision Making for Food Waste Composting Management." *Journal of Advanced Research in Technology and Innovation Management* 2, no. 1 (2022): 1-12.

[22] Siam, Mohd Firdaus Mohd, Ahmad Azad Ab Rashid, Nurulhana Borhan, and Mohd Khairul Alhapiz Ibrahim. "Distracted driving while doing mobile phone conversation: A driving simulator study." *Journal of Advanced Research Design* 56, no. 1 (2019): 1-9.

[23] Zamani, Hadi, and Muhamad Kamal Mohammed Amin. "Classification of phishing websites using machine learning techniques." *Journal of Advanced Research in Applied Sciences and Engineering Technology* 5, no. 2 (2016): 12-19.

[24] Panwar, Surya Nandan, Saliya Goyal, and Prafulla Bafna. "Analytical Study of Starbucks Using Clustering." In *International Conference on Hybrid Intelligent Systems*, pp. 1013-1021. Cham: Springer Nature Switzerland, 2022. https://doi.org/10.1007/978-3-031-27409-1_93

[25] Kaartinen, Niina E., Minna E. Similä, Noora Kanerva, Liisa M. Valsta, Kennet Harald, and Satu Männistö. "Naturally occurring and added sugar in relation to macronutrient intake and food consumption: results from a population-based study in adults." *Journal of nutritional science* 6 (2017): e7. https://doi.org/10.1017/jns.2017.3

[26] Popkin, Barry M., Lawrence E. Armstrong, George M. Bray, Benjamin Caballero, Balz Frei, and Walter C. Willett. "A new proposed guidance system for beverage consumption in the United States." *The American journal of clinical nutrition* 83, no. 3 (2006): 529-542. https://doi.org/10.1093/ajcn.83.3.529