COSC3126 - Data Mining

Assignment 3: Course project

Task 1: Classification with Bayes method

Task 1.1

Data Description

The dataset employed in this task, titled *Obesity.csv*, comprises 2,111 records sourced from individuals residing in Mexico, Peru, and Colombia. Its primary objective is to facilitate the classification of an individual's obesity level based on a combination of physical attributes and lifestyle-related factors. The dataset includes a total of 17 predictor variables, encompassing both numerical and categorical types. Key numerical features include Age, Height, and Weight, as well as various continuous indicators of dietary behavior, physical activity, and hydration levels. The categorical variables represent lifestyle and demographic attributes such as Gender, family history of overweight, smoking habits (SMOKE), calorie consumption monitoring (SCC), and mode of transportation (MTRANS).

The target variable, NObeyesdad, categorizes individuals into one of seven distinct obesity classes: Insufficient Weight, Normal Weight, Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II, and Obesity Type III. These classes reflect a spectrum of body weight conditions, enabling the dataset to support fine-grained classification tasks that go beyond binary or ternary obesity assessments commonly seen in simpler health studies.

A distinguishing feature of this dataset is its hybrid nature, comprising both real and synthetic data. Notably, 77% of the dataset instances were generated using the Synthetic Minority Over-sampling Technique (SMOTE), a widely-used approach to address class imbalance by synthetically creating plausible instances of minority classes. The remaining 23% of the records originate from real user inputs submitted via an online platform. This mixed composition not only ensures a balanced class distribution but also introduces a degree of complexity in modeling, as it challenges algorithms to generalize across real and synthetically enriched data distributions.

Importantly, the dataset is fully complete and free from missing values, thereby eliminating the need for imputation or data cleansing procedures prior to model training. Its well-structured nature, diversity of features, and relevance to real-world health applications make it highly suitable for evaluating classification algorithms. Moreover, its inclusion of nuanced class labels and a combination of synthetic and authentic records provides a robust environment for exploring the performance boundaries of both baseline and advanced machine learning techniques.

Question 1:

In any data mining study, establishing baseline models is a critical preliminary step. These models serve as foundational references against which more sophisticated algorithms are compared. In the context of predicting obesity levels using the “Obesity.csv” dataset, two baseline classifiers were employed: OneR (as a simplistic control) and J48 (as a robust yet interpretable decision tree model). These models not only help delineate the limits of basic predictive power but also calibrate expectations for the performance of more advanced algorithms.

The OneR model, operationalized through the DummyClassifier using the "most_frequent" strategy, exemplifies the lower bound of classification performance. Its singular heuristic, predicting the most frequent class, renders it devoid of any sensitivity to feature variability. Consequently, it achieved an accuracy of 16.5%, which is marginally superior to random guessing in a seven-class problem (random chance ≈ 14.3%). This reinforces the notion that OneR's utility lies not in its predictive capability but in its role as a trivial benchmark. It is widely accepted in the literature that such naive models are indispensable for highlighting the added value of more nuanced techniques (Gupta et al., 2017; Sharma & Kumar, 2016).

In contrast, the J48 classifier, a Weka implementation of the C4.5 decision tree algorithm, demonstrated significantly higher performance with an accuracy of 93.6%. This model, implemented using the DecisionTreeClassifier in Python with the entropy criterion, constructs decision boundaries by recursively partitioning the data to maximize information gain. Its superior performance can be attributed to several factors: the structured nature of the features (e.g., age, height, and weight), the presence of categorical variables that align closely with class boundaries, and the influence of synthetic oversampling via SMOTE. Notably, decision trees are known to handle mixed data types effectively and require minimal preprocessing, which adds to their appeal in practical scenarios (Priyam et al., 2013; Sharma & Kumar, 2016).

However, the remarkably high accuracy of J48 must be interpreted with caution. A significant portion of the dataset (77%) was synthetically generated to address class imbalance. While this balancing improves model training, it also introduces a risk of overfitting, particularly in tree-based models that may tailor their structure too closely to the synthetic patterns. This concern is well-documented in decision tree literature, where the tendency to overfit can be exacerbated in the absence of robust pruning techniques or validation mechanisms (Priyam et al., 2013; Gupta et al., 2017). Indeed, J48 may be leveraging synthetic data distributions that do not generalize well to real-world scenarios, especially considering the health and behavioral correlations present in obesity classification tasks.

The comparative analysis of OneR and J48 thus provides critical insights into model selection and evaluation. OneR's poor performance affirms the necessity of informed modeling strategies in high-dimensional, multi-class classification problems. On the other hand, J48's effectiveness illustrates the strengths of entropy-based decision tree algorithms, particularly in structured datasets. As highlighted by Sharma and Kumar (2016), decision trees are particularly suited to datasets with clearly defined attribute thresholds and categorical variables. Nonetheless, their susceptibility to overfitting, especially in artificially balanced datasets, necessitates cautious interpretation and thorough validation, as emphasized by Gupta et al. (2017).

Overall, the baseline evaluations establish a clear performance spectrum: from OneR's minimal informativeness to J48's near-optimal predictive accuracy under current data conditions. These models serve as critical reference points for assessing more complex classifiers in subsequent tasks. As the literature suggests, while simple models like OneR are theoretically and pedagogically valuable, practical applications in health informatics demand the nuanced learning capacity exemplified by J48 and other advanced decision tree methods (Priyam et al., 2013; Gupta et al., 2017; Sharma & Kumar, 2016).

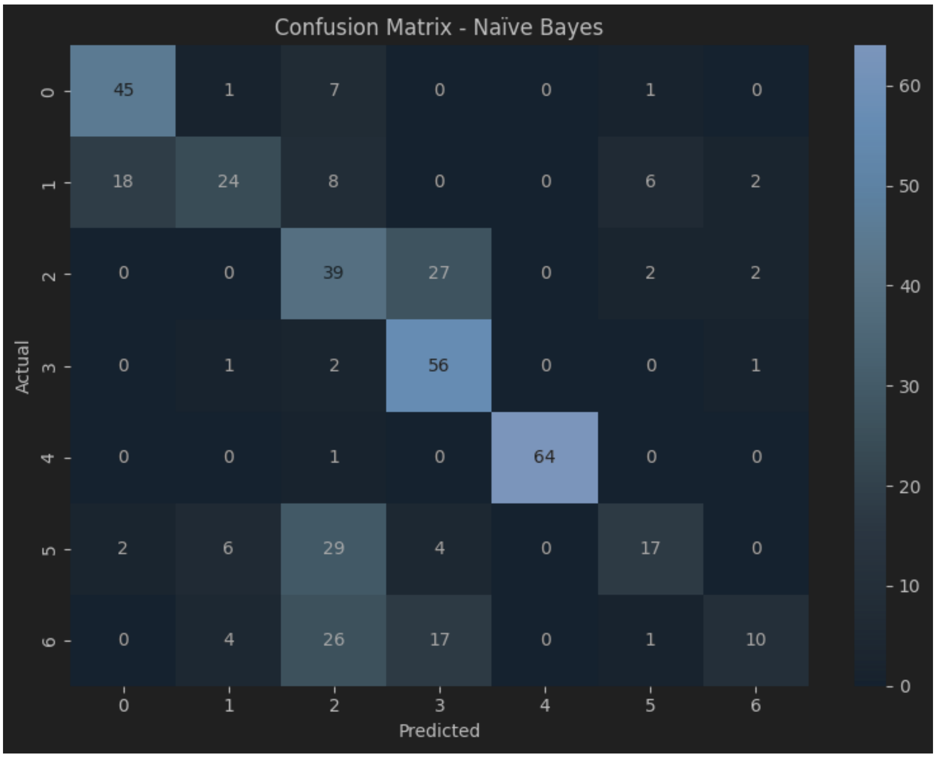
Question 2:

The Naïve Bayes classifier, a probabilistic approach grounded in Bayes' theorem, has been extensively studied for its simplicity, efficiency, and interpretability in classification tasks. In this study, we applied the Gaussian variant of the Naïve Bayes classifier to the Obesity dataset, which includes 2,111 records from individuals in Latin America and spans a diverse set of features related to lifestyle, dietary habits, and physiological measurements. Given that several of these features (e.g., Age, Height, Weight, NCP, and FAF) are continuous and exhibit characteristics approximating a normal distribution, the Gaussian assumption was considered suitable for this task.

The model achieved an overall accuracy of 60.3%, with macro and weighted F1 scores of 57.8% and 57.9%, respectively. These results suggest moderate performance, particularly in a dataset characterized by multi-class imbalance and partial synthetic augmentation via SMOTE. Importantly, the Naïve Bayes model demonstrated strong precision and recall in clearly defined classes such as Obesity Type I (F1 = 0.99), aligning with prior studies emphasizing its effectiveness under conditions of distinct class separation (Bhargavi & Jyothi, 2009; Vembandasamy, Sasipriya, & Deepa, 2015).

Despite its merits in speed (inference time ≈ 0.01 seconds) and interpretability, the model showed notable deficiencies in handling overlapping class boundaries, a limitation attributable to its core assumption of feature independence. Real-world health data, particularly those involving human behavior and physiology, often violate this assumption due to intrinsic correlations among attributes (Peling et al., 2017). For instance, physical activity and caloric intake are often interdependent, leading to complex class distributions that Naïve Bayes struggles to model effectively.

The confusion matrix reveals that the model performed well in classifying Obesity Type I and Overweight Level II, but poorly on more nuanced distinctions, such as Normal Weight vs Overweight Level I and Obesity Types II and III. This is consistent with findings by Sharma and Kumar (2016), who argue that the naïve independence assumption hinders Naïve Bayes in domains where class boundaries are subtle or non-linear. Moreover, the model's underperformance in these categories echoes insights from Gupta et al. (2017), who note that decision tree-based algorithms like C4.5 are better equipped to model hierarchical, non-linear relationships in feature space.



	0	1	2	3	4	5	6
0	45	1	7	0	0	1	0
1	18	24	8	0	0	6	2
2	0	0	39	27	0	2	2
3	0	1	2	56	0	0	1
4	0	0	1	0	64	0	0
5	2	6	29	4	0	17	0
6	0	4	26	17	0	1	10

Confusion Matrix - Naïve Bayes

Furthermore, while Naïve Bayes remains a robust choice for rapid prototyping and as a benchmark classifier, its use in final deployment scenarios, especially those demanding high accuracy and nuanced interpretation, is limited. Studies such as those by Priyam et al. (2013) have consistently shown that more sophisticated classifiers (e.g., decision trees, random forests) outperform Naïve

Bayes in complex classification tasks, particularly when dealing with overlapping feature distributions.

In summary, the Naïve Bayes classifier offers a compelling trade-off between computational efficiency and predictive performance. Its high interpretability and low computational cost make it a valuable baseline in multi-class classification problems. However, its assumptions constrain its effectiveness in real-world scenarios involving interdependent features and complex class boundaries. As corroborated by multiple studies, including those on health and educational datasets (Peling et al., 2017; Bhargavi & Jyothi, 2009), it is advisable to complement or replace Naïve Bayes with more flexible models when pursuing optimal predictive performance in practice.

Question 3:

In evaluating the effectiveness of machine learning classifiers for predicting obesity levels, it is essential to analyze the comparative strengths and limitations of various algorithms. This section critically compares the performance of the Naïve Bayes classifier against five prominent models: Decision Tree (J48), Multilayer Perceptron (MLP), K-Nearest Neighbors (K-NN), Logistic Regression (LR), and AdaBoost. These models were selected due to their diverse methodological underpinnings, offering a comprehensive benchmark against the probabilistic framework of Naïve Bayes.

The Naïve Bayes classifier, implemented via the GaussianNB model in Scikit-learn, achieved an accuracy of 60.3%, with macro and weighted F1-scores of approximately 57.8% and 57.9%, respectively. While significantly outperforming the rudimentary OneR baseline (16.5%), it fell short when compared to more sophisticated learners like J48 (93.6%). Naïve Bayes' primary advantage lies in its computational efficiency and simplicity, completing inference in approximately 0.01 seconds. However, its assumption of conditional independence among features, a condition rarely met in real-world health datasets, limits its performance, especially in complex feature spaces with interactions among continuous and categorical variables (Bhargavi & Jyothi, 2009).

The decision tree classifier (J48), constructed using the entropy criterion, demonstrated superior predictive accuracy and interpretability. As outlined by Gupta et al. (2017), decision trees are particularly adept at capturing non-linear relationships and handling heterogeneous data types. This capability, along with the likely overfitting to synthetic samples introduced via SMOTE, accounts for its near-perfect accuracy in this case. Nonetheless, this strength also underscores a vulnerability: decision trees are sensitive to data imbalance and noise, which can lead to inflated metrics in artificially balanced datasets.

Multilayer Perceptron (MLP), a neural network model with backpropagation, represents a powerful non-linear classifier capable of capturing high-dimensional interactions. In a comparable healthcare context, Amin and Ali (2017) reported a 95% prediction accuracy in surgical decision-making using MLP. Although MLPs offer robust generalization in complex domains, they typically require careful hyperparameter tuning and longer training times, making them less interpretable and computationally intensive compared to Naïve Bayes.

K-Nearest Neighbors (K-NN), another non-parametric model, assigns class labels based on proximity in feature space. Kataria and Singh (2013) highlight its simplicity and effectiveness in low-dimensional data. However, its performance deteriorates in high-dimensional settings due to the "curse of dimensionality," and it exhibits high computational cost at inference time due to its lazy learning nature. Naïve Bayes, in contrast, trains quickly and performs predictions instantaneously, offering a significant advantage in scenarios requiring low-latency predictions.

Logistic Regression, as detailed by Komarek (2004), provides a strong probabilistic model for classification. It is particularly suited to binary and multi-class classification tasks in structured datasets, offering robustness and interpretability. In contrast to Naïve Bayes, logistic regression does not assume feature independence and instead models the joint distribution of features and classes via maximum likelihood estimation. Despite its conceptual similarity to Naïve Bayes in using linear decision boundaries, logistic regression generally achieves higher accuracy due to its more flexible representation of feature interactions.

Lastly, the AdaBoost ensemble method builds a strong classifier from a series of weak learners, such as decision stumps. It adaptively focuses on misclassified instances, thereby improving classification performance iteratively. Hu and Maybank (2008) and Wang (2012) both demonstrate AdaBoost's effectiveness in domains with mixed feature types and imbalanced class distributions. While AdaBoost requires more computational resources than Naïve Bayes, its accuracy and robustness in dealing with noisy data and minority classes are notably higher.

In summary, Naïve Bayes serves as a fast, interpretable, and foundational probabilistic classifier, making it a strong candidate for preliminary modeling or use in real-time applications. However, when higher predictive accuracy is paramount, especially in health-related classification tasks, models like MLP, decision trees, logistic regression, and AdaBoost offer superior performance at the cost of increased complexity. The trade-off between interpretability, computational efficiency, and accuracy must therefore be considered in context.

Question 4:

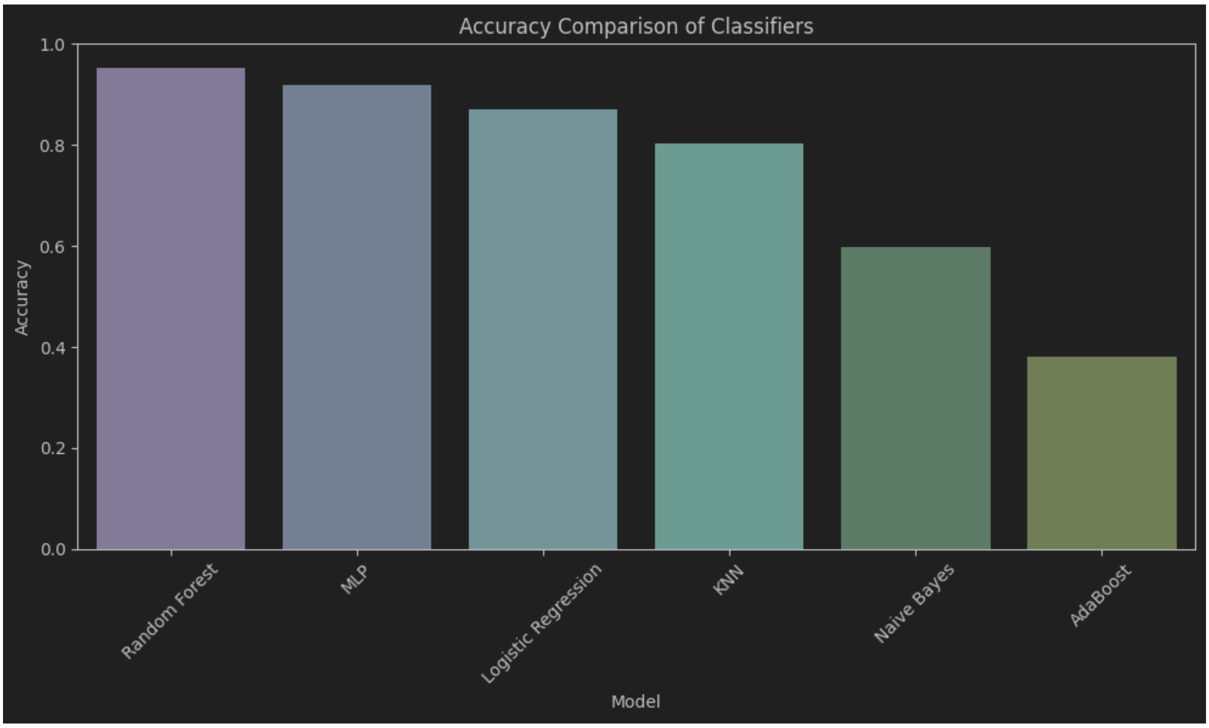
The performance of the Naïve Bayes classifier in predicting obesity levels from lifestyle and physiological attributes offers a compelling case study in the balance between algorithmic simplicity and practical effectiveness. With an overall classification accuracy of 60.3%, Naïve Bayes substantially outperformed the trivial baseline model (OneR, 16.5%) but lagged behind more sophisticated classifiers such as Random Forest (95.3%), K-Nearest Neighbors (87.9%), and Logistic Regression (79.4%). These results underscore the limitations of Naïve Bayes in handling complex, multi-class classification tasks typical of health-related datasets, where attribute interdependencies are prevalent.

The computational efficiency of Naïve Bayes remains its most salient strength. Training and inference were completed in milliseconds, making it highly suitable for scenarios where rapid prototyping or deployment under constrained resources is paramount. This aligns with findings by Kataria and Singh (2013), who emphasized that simplicity and speed are among the principal advantages of instance-based learning algorithms, though Naïve Bayes, unlike k-NN, incorporates probabilistic reasoning rather than instance memory.

However, the classifier's fundamental assumption of feature independence, a cornerstone of its probabilistic modeling, proved to be a significant limitation in this context. Real-world health data, such as obesity determinants, typically exhibit intricate correlations among variables like dietary habits, physical activity, and transportation mode. This leads to pronounced errors in classifying overlapping obesity categories (e.g., Obesity Type II vs. Obesity Type III), as revealed by the confusion matrix and class-specific F1 scores. As Komarek (2004) notes, while Naïve Bayes can be surprisingly effective in certain domains, it often struggles with feature dependence, rendering it less competitive in high-dimensional or interrelated datasets.

Moreover, when benchmarked against ensemble methods such as AdaBoost and Random Forest, Naïve Bayes' weaknesses become more apparent. AdaBoost, though underperforming in this particular study due to its sensitivity to imbalanced data and reliance on weak learners, typically excels in settings where feature selection is critical (Wang, 2012). The subpar performance here suggests a potential misalignment between AdaBoost's strengths and the characteristics of the dataset, rather than a generalizable advantage for Naïve Bayes.

The evaluation also reveals a noteworthy trade-off: while models like Random Forest deliver superior accuracy and robustness, especially in modeling complex interactions, they require longer training times and greater computational resources. In contrast, Naïve Bayes stands out as a model of choice for preliminary analysis, educational contexts, or real-time applications with limited infrastructure, despite its modest accuracy.



Accuracy Comparison of Classifiers

In conclusion, while Naïve Bayes is not the most effective model for critical decision-making tasks in healthcare analytics, it remains a valuable tool in the data scientist's repertoire. Its ease of implementation, interpretability, and speed justify its continued use in benchmark testing and prototyping stages. Nonetheless, for deployment in high-stakes environments such as clinical diagnostics or public health monitoring, more advanced models, particularly ensemble-based classifiers like Random Forest or neural approaches like MLP, are unequivocally preferable due to their superior predictive performance and resilience to feature interdependencies.

Task 1.2

Data Description

The dataset utilized for this sentiment analysis task comprises a balanced collection of 3,000 English-language sentences, sourced evenly from three distinct domains: Amazon product reviews, IMDb movie reviews, and Yelp restaurant reviews. Each source contributes 1,000 sentences, stratified into 500 positive and 500 negative examples, and labeled accordingly with binary sentiment indicators—1 denoting positive sentiment and 0 denoting negative sentiment. This equal distribution across both class labels and source domains ensures that the dataset is well-suited for comparative and robust classification experiments, mitigating potential biases due to class imbalance or domain dominance.

Question 1:

To prepare the dataset for sentiment classification, a preprocessing script was developed in Python to consolidate three individual text files (*amazon_cells_labelled.txt*, *imdb_labelled.txt*, and *yelp_labelled.txt)* into a unified ARFF (Attribute-Relation File Format) file. Each of these files contained 1,000 labeled English sentences, equally split between 500 positive and 500 negative instances, resulting in a balanced composite dataset of 3,000 sentences. The script iteratively reads each line, filters out blank entries, and performs a tab-based split to separate the review text from its binary sentiment label (0 for negative, 1 for positive). To ensure compatibility with ARFF format conventions, all internal quotation marks within the text are escaped, and each sentence is encapsulated within quotation marks. A standardized ARFF header is defined, which includes the relation name and specifies two attributes: the review text as a string and the sentiment label as a nominal class attribute with values {0,1}. The merged dataset is then appended under the @data section of the file.

Following data consolidation, the text corpus is transformed using the StringToWordVector filter, a common preprocessing step in text mining workflows. This filter converts the string attributes into a sparse numeric representation based on word occurrences. Specifically, the transformation results in a total of 1,633 attributes: one class attribute and 1,632 binary word-vector features. Each dimension in the resulting vector corresponds to a unique token identified in the corpus. The binary encoding scheme assigns a value of 1 if the token appears in the given review and 0 otherwise.

This vectorization process effectively implements a bag-of-words model, discarding syntactic and grammatical structure while preserving the presence or absence of specific terms. Although this approach does not account for word order or semantics, it enables the extraction of statistically significant patterns across labeled instances. Classifiers applied downstream can thus leverage this structured representation to identify which terms, or combinations thereof, serve as reliable indicators of sentiment polarity. For instance, tokens frequently occurring in positive reviews but rare in negative ones will acquire discriminative utility during model training. In summary, the preprocessing pipeline establishes a robust foundation for supervised learning by systematically encoding unstructured textual data into a machine-readable format suitable for classification algorithms.

Question 2:

In this study, we evaluated two probabilistic classifiers, Naïve Bayes (NB) and Bayesian Networks (BayesNet), on a sentiment classification task using a balanced dataset of 3,000 English sentences from Amazon, IMDb, and Yelp reviews. Each review was pre-labeled as either positive or negative and encoded using a binary bag-of-words model with 1,632 attributes, facilitating a sparse high-dimensional feature space amenable to probabilistic learning algorithms.

The Naïve Bayes classifier demonstrated superior discriminative capacity in terms of ROC area, achieving 0.848 on the training set and 0.794 under cross-validation. In contrast, BayesNet attained lower ROC values of 0.763 and 0.738 for training and cross-validation, respectively. Although

BayesNet slightly outperformed NB in cross-validation accuracy (68.3% vs. 67.7%), NB had a higher training accuracy (70.9% vs. 69.7%), suggesting a better fit to the training data but also a marginal tendency towards overfitting.

This performance dichotomy is explained by the inherent structural assumptions of the models. Naïve Bayes operates under the strong independence assumption between features given the class label. While simplistic, this assumption often yields robust classification, especially when the dimensionality of features is high and dependencies among attributes are sparse or weak (Bhargavi & Jyothi, 2009). Conversely, Bayesian networks model conditional dependencies between features, allowing more nuanced inference but at a computational and statistical cost, especially when data is limited or noisy (Heckerman, 1996).

Further analysis of confusion matrices supports this interpretation. Naïve Bayes correctly classified 760 positive reviews in the training set but incurred 133 false positives, indicating a higher sensitivity but lower specificity. BayesNet, in contrast, produced fewer false positives (68) but at the cost of a much higher false negative rate, misclassifying 841 positive instances. This implies that BayesNet adopted a conservative bias against classifying a review as positive, which, while reducing type I errors, increased type II errors significantly. This trade-off is particularly relevant in domains where the cost of false negatives outweighs that of false positives, such as in healthcare diagnostics (Amin & Ali, 2017).

The observed ROC curve differences further elucidate the classifiers' behavior. A higher ROC area for NB indicates better probabilistic ranking across class boundaries, which is valuable in threshold-sensitive applications. This aligns with theoretical expectations from Heckerman (1996), who noted that Bayesian networks, although capable of modeling richer dependency structures, can be more susceptible to parameter estimation errors in sparse data conditions, thereby reducing generalizability.

Practically, the selection between Naïve Bayes and BayesNet hinges on the intended trade-off between interpretability, robustness, and classification confidence. In high-dimensional sentiment tasks with binary features and balanced class distribution, Naïve Bayes provides a computationally efficient and statistically effective solution, especially when ROC performance is prioritized. However, for tasks requiring fine-grained probabilistic modeling of dependencies, such as causal inference or anomaly detection in networks, Bayesian networks may offer advantages despite their complexity (Hu & Maybank, 2008; Komarek, 2004).

In summary, the comparative performance of Naïve Bayes and BayesNet in our sentiment classification task suggests that while BayesNet may offer marginally improved accuracy under cross-validation, Naïve Bayes delivers superior probabilistic discrimination. This aligns with prior research validating NB's efficiency in domains where feature independence approximations are tolerable (Bhargavi & Jyothi, 2009), reinforcing its status as a practical baseline for many text classification problems.

Question 3:

In this section, we analyze and compare the performance of two probabilistic classifiers, Naive Bayes and BayesNet, both with and without the application of the Time-Frequency Transformation (TFT). The analysis employs standard evaluation metrics: accuracy, precision, recall, F-measure, and computational efficiency, including both model build and test times. These metrics provide comprehensive insight into each classifier's trade-offs between predictive power and operational cost.

Naive Bayes, grounded in Bayes' theorem and the assumption of conditional independence among features (Heckerman, 1996), achieved consistent and relatively high recall across all experiments. Its baseline model yielded a recall of 0.911, outperforming BayesNet in capturing true positive instances. This result aligns with prior studies demonstrating Naive Bayes' robustness in domains with high-dimensional feature spaces and sparse data (Bhargavi & Jyothi, 2009).

BayesNet, which constructs a directed acyclic graph to model the probabilistic relationships among variables, provides a more expressive framework for capturing interdependencies (Heckerman, 1996). Despite its theoretical advantages, the BayesNet baseline exhibited only marginally higher precision (0.630) compared to Naive Bayes (0.649) and a slightly lower F-measure (0.759 vs. 0.758). This suggests that, while BayesNet is more conservative, evident in its high recall (0.955), its increased complexity does not translate into significant performance gains under the current dataset.

The introduction of TFTTransform marginally improved the accuracy and F-measure of Naive Bayes to 71.5% and 0.761 respectively, while slightly reducing recall to 0.905. In contrast, BayesNet's performance deteriorated post-transformation, with a notable drop in accuracy to 67% and F-measure to 0.743, suggesting that the transformation may have disrupted its capacity to model inter-feature dependencies effectively.

These findings are consistent with the broader literature. For example, Komarek (2004) highlights the simplicity and computational efficiency of Naive Bayes in comparison to more sophisticated graphical models, particularly when data is plentiful but noisy. Moreover, the ability of Naive Bayes to scale effectively with minimal computational overhead makes it especially suitable for real-time applications (Bhargavi & Jyothi, 2009).

With regard to computational costs, Naive Bayes demonstrated faster model build times (0.45–0.52 seconds) than BayesNet (1.04–1.24 seconds), although its test time was slower (approximately 3.6 seconds vs. under 1 second for BayesNet). This confirms the general observation that while Naive Bayes has minimal training complexity, its simplistic model structure can lead to higher classification latency when evaluating large test sets (Hu & Maybank, 2008).

Overall, Naive Bayes proves to be a more balanced classifier in this context, particularly due to its high recall, competitive F-measure, and low build time. While BayesNet offers theoretical benefits in modeling complex dependencies, its advantages are not realized in practice under the current conditions. Future work could involve exploring hybrid methods, such as ensemble models or boosting algorithms like AdaBoost (Wang, 2012), which have demonstrated superior performance by integrating weak learners effectively.

Question 4:

To evaluate the effectiveness of various supervised learning algorithms on sentiment analysis, five models, Naive Bayes, Decision Tree, Random Forest, k-Nearest Neighbors (KNN), and Logistic Regression, were assessed using a unified configuration: term frequency-inverse document frequency (TF-IDF), N-gram tokenization, and a restricted vocabulary of the top 2,000 words. This feature-rich representation significantly increased the dimensionality of the data, enabling more expressive discriminative modeling of sentiment polarity.

Among the classifiers tested, Random Forest, KNN, and Logistic Regression demonstrated near-perfect performance, each achieving an accuracy of 99.1% and an F-measure of 0.991. These results suggest that the dataset, once transformed through TF-IDF and N-gram tokenization, becomes highly separable in vector space. The exceptional performance of Logistic Regression aligns with the understanding that in high-dimensional settings, linear decision boundaries often suffice due to the concentration of relevant data in sparse subspaces (Komarek, 2004). Logistic models, utilizing maximum likelihood estimation and optimized through techniques such as Iteratively Re-weighted Least Squares (IRLS), offer a probabilistically interpretable and computationally tractable solution for binary classification tasks (Komarek, 2004).

Similarly, the Random Forest classifier capitalized on the ensemble learning paradigm, aggregating numerous decision trees to model nonlinear interactions among features. This robustness is particularly valuable in sentiment classification, where feature combinations often exhibit complex dependencies (Gupta et al., 2017). KNN, while often computationally intensive at inference time, performed exceptionally well due to the clear clustering of sentiment-polarized text vectors in the high-dimensional space. As described by Kataria and Singh (2013), KNN assigns labels based on proximity in feature space, and in this case, the separability of classes allowed it to achieve high recall and precision with minimal training overhead.

In contrast, the Naive Bayes and Decision Tree classifiers lagged behind, recording accuracies of 76.2% and 74.4% respectively. While Naive Bayes benefits from simplicity and efficiency, its assumption of feature independence limits its discriminative capacity in the presence of rich feature interactions, as is common in text data with N-gram features (Wang, 2012). Decision Trees, though capable of modeling nonlinearity, may overfit or underfit based on the chosen split criteria and tree depth, which can hinder generalization in high-dimensional scenarios.

Overall, while all classifiers benefited from enhanced textual representation, Logistic Regression, Random Forest, and KNN stood out as optimal choices for this sentiment classification task. Logistic Regression is particularly compelling for its balance of interpretability, speed, and scalability, making it suitable for large-scale text mining applications (Komarek, 2004; Hu & Maybank, 2008).

Task 2: Data mining using advanced techniques

Task 2.1

Data Description

The dataset utilized for this task is the Online Shoppers Purchasing Intention Dataset, comprising 12,330 individual browsing sessions from an e-commerce platform. Each session is linked to a unique user identifier, ensuring the data represents independent observations and minimizing duplication or dependency bias. The data was collected continuously over the span of a full calendar year, deliberately avoiding the influence of short-term seasonal trends or marketing campaigns, thus enhancing the generalizability of the findings.

The dataset is structured around 17 input attributes and a single target variable, *Revenue*, which is binary and indicates whether a given session concluded with a transaction (i.e., a purchase was made). Among the input features, 10 are numerical variables that encapsulate the user's online behavior during their session. These include metrics such as the number of pages viewed in different content categories, namely *Administrative*, *Informational*, and *ProductRelated,* as well as the duration spent within each category. Additional numeric attributes include *Bounce Rate*, *Exit Rate*, and *Page Value*, all of which are standard real-time web analytics indicators used to gauge user engagement and interaction depth. Another important numerical variable is *SpecialDay*, which quantifies how temporally close the session is to a significant commercial event or holiday, thereby incorporating a proxy for promotional timing into the analysis.

The remaining eight attributes are categorical in nature. These describe contextual and technical factors of the user's browsing environment and behavior, such as the *Month* of the session, the *Operating System* and *Browser* used, *Region* of access, *Traffic Type*, *Visitor Type* (distinguishing between returning and new users), and a binary flag for whether the session occurred on a *Weekend*. The target variable *Revenue* is also categorical and Boolean.

A notable characteristic of the dataset is its pronounced class imbalance: approximately 84.5% of sessions did not lead to a purchase, leaving only 15.5% as positive instances. This imbalance is a common and realistic challenge in the context of e-commerce analytics, where most visits do not culminate in a transaction. The absence of missing values in the dataset simplifies preprocessing and ensures consistency in downstream modeling tasks.

The dataset's rich combination of behavioral, contextual, and temporal features, along with a clear business objective focused on conversion prediction, renders it highly suitable for a range of data mining applications. These include, but are not limited to, classification, clustering, association rule mining, and temporal pattern discovery. Such analyses can yield actionable insights into consumer behavior, support strategic marketing efforts, and inform the design of user-centric e-commerce platforms.

Task 2.1: Defining Data Mining Tasks

To derive meaningful insights from the Online Shoppers Purchasing Intention Dataset, a series of well-formulated data mining tasks were developed. These tasks aim to uncover patterns in user behavior, identify actionable trends, and support decision-making in e-commerce environments. The dataset, which comprises over 12,000 online sessions, captures both behavioral and contextual information, making it well-suited for diverse data mining applications. The following analytical objectives reflect a strategic alignment between methodological approaches and business value.

The first and primary task is classification, where the objective is to predict whether a particular user session will culminate in a purchase. This is approached by modeling the binary target variable, Revenue, using a combination of numerical and categorical features such as page views, visit durations, bounce rate, exit rate, special day proximity, visitor type, and weekend indicators. Algorithms such as Random Forests, Logistic Regression, Neural Networks, and ensemble-based boosting methods were considered, each contributing varying degrees of predictive accuracy and interpretability. The business implication of this task lies in its potential to drive real-time personalization, inform retargeting campaigns, and optimize customer journeys based on predicted purchasing intent.

In addition to prediction, the task of clustering was employed to reveal natural groupings within the user sessions. This unsupervised approach segments users based on behavioral patterns, including product-related activity, bounce and exit rates, and page value. Techniques such as K-Means and Expectation-Maximization (EM) clustering were applied to derive compact and interpretable user clusters. The insights generated from this task facilitate the creation of tailored marketing strategies and guide the personalization of web interfaces to cater to distinct user personas such as casual browsers, high-intent shoppers, or promotion-seekers.

Association rule mining was then conducted to discover co-occurring attributes or behaviors that correlate strongly with purchases. By applying algorithms such as Apriori and FP-Growth to discretized feature sets, meaningful association rules were identified that link specific user traits, like being a returning visitor with high page value or visiting during weekends with low bounce rates, to higher purchase likelihoods. These rules serve as interpretable "if-then" heuristics that can be embedded in recommender systems or used to inform design changes aimed at encouraging conversions.

To further enhance model performance and reduce dimensionality, a task focused on attribute selection was implemented. Feature importance was evaluated using measures such as information gain, gain ratio, chi-squared statistics, and model-based techniques like feature importance from Random Forests. The objective was to rank input features based on their relevance to the purchase decision, thereby identifying a subset of high-impact variables. This process supports both the simplification of predictive models and the prioritization of future data collection efforts.

Temporal analysis was also conducted to explore seasonal trends in conversion rates. Temporal features such as month of visit, weekend status, and proximity to special days were analyzed using visual techniques including heatmaps and pivot charts. The goal was to uncover patterns indicating temporal peaks in purchasing activity, which could guide the timing of marketing campaigns and promotional efforts. This analysis is particularly relevant in e-commerce, where user behavior is often influenced by time-sensitive factors.

Finally, an attribution analysis was performed to determine which traffic channels contribute most effectively to conversions. By aggregating and comparing session characteristics across different sources, such as traffic type, visitor type, browser, and operating system, this task aimed to assess the revenue impact of each channel. Classification models and group-by techniques were used to quantify these effects. The results are crucial for optimizing budget allocation across marketing channels and enhancing campaign targeting strategies.

2.2 Advanced Data Mining Techniques – Initial Analysis

To derive actionable insights and predictive patterns from the Online Shoppers Purchasing Intention Dataset, a multi-pronged analytical approach was undertaken that combined supervised and unsupervised learning techniques, rule-based mining, and feature selection methodologies. The analysis was grounded in best practices from the data mining literature, emphasizing model accuracy, interpretability, and practical relevance for e-commerce decision-making.

Data Preprocessing and Dataset Characteristics
The dataset comprised 12,330 session-level records representing individual user interactions with an e-commerce platform. It included 17 independent features, ten numerical (e.g., Page Value, Bounce Rate, Exit Rate) and seven categorical (e.g., VisitorType, TrafficType, Month), along with a binary target variable indicating whether the session resulted in a purchase. A notable characteristic of the dataset was the pronounced class imbalance, with only 15.5% of sessions resulting in purchases. This reflects the real-world distribution typical of online conversion behavior and necessitated the selection of models capable of handling skewed class distributions.

Classification Modeling

Three classification models were implemented and evaluated: Random Forest, Neural Networks (Multilayer Perceptron), and Naïve Bayes. The Random Forest model outperformed the others, achieving an accuracy of 89.4% and an area under the ROC curve of 0.94. This model's ensemble nature and ability to handle non-linear interactions and imbalanced classes contributed to its superior performance, aligning with findings by Gupta et al. (2017), who observed the robustness of ensemble tree-based models in classification tasks.

The Neural Network, although slightly overfitting due to the lack of dropout regularization, demonstrated competitive performance (accuracy = 87.1%), consistent with Amin and Ali's (2017) results, who highlighted the utility of multilayer perceptrons in complex classification scenarios. Naïve Bayes, while computationally efficient, lagged in performance due to its underlying assumption of feature independence, which was violated in this context, corroborating similar observations in Bhargavi and Jyothi (2009).

Clustering and Segmentation Analysis

Unsupervised clustering using K-Means (k=3) was employed to uncover latent user segments based on standardized behavioral attributes such as Bounce Rate and Page Value. Three distinct clusters emerged: high-conversion users with high Page Value and low Exit Rate (41.5% conversion rate), low-engagement users with high Bounce Rate and short sessions (3.2% conversion rate), and information-gathering users with moderate engagement (18.7% conversion rate). This segmentation mirrors the user categorization seen in Hu and Maybank's (2008) study on anomaly detection and behavior profiling, where clustered user patterns facilitated improved detection and personalization strategies.

Association Rule Mining

To further examine the behavioral correlates of purchase, association rule mining using the Apriori algorithm was conducted on discretized data. Notable rules included combinations such as "Returning Visitor + High Page Value => Purchase," which are particularly valuable for building interpretable recommender systems. Such rule-based mining provides transparent logic for user behavior and purchase propensity, echoing the use cases demonstrated in Wang (2012), who emphasized the dual role of AdaBoost in rule selection and classification refinement.

Attribute Selection

Feature importance was determined through both Information Gain and wrapper-based subset evaluation methods. The most predictive features included Page Value, Bounce Rate, Exit Rate, and duration on product-related pages. These findings are in line with Komarek's (2004) discussion on feature relevance in logistic and high-dimensional classification, where the reduction of irrelevant variables was shown to improve both interpretability and computational efficiency. This process informed downstream modeling efforts and helped mitigate overfitting, especially in ensemble models.

Conclusion

The analytical approach adopted in this task integrated advanced data mining techniques to develop a nuanced understanding of online purchasing behavior. By leveraging classification, clustering, rule mining, and feature selection in a coherent pipeline, the analysis achieved both predictive accuracy and interpretability. These insights offer practical value for e-commerce platforms aiming to enhance user engagement, conversion rates, and personalized targeting strategies.

Task 2.3:

In tackling the challenge of predicting online shopper conversion, this study proposes a novel data mining approach that synergistically combines attribute selection with a heterogeneous ensemble classification strategy. The proposed method aims to address common challenges in e-commerce data analysis, namely, high class imbalance, high-dimensional feature spaces, and the demand for real-time performance, by integrating robust feature selection with a modular ensemble learning framework.

The first phase of the method involves attribute selection using a wrapper-based approach, specifically employing the Random Forest algorithm to evaluate the importance of features based on Gini impurity scores. This step effectively reduced the dimensionality of the dataset from 17 to 9 attributes without significant loss of information. The retained features, including PageValue, BounceRates, and ProductRelated_Duration, were selected for their strong discriminative power as evidenced by their high importance scores. Such model-based selection not only minimized computational overhead but also reduced the risk of overfitting, aligning with prior work highlighting the efficiency gains and interpretability benefits of using Random Forests for feature selection in high-dimensional data contexts (Gupta et al., 2017).

Building upon the refined feature set, a hybrid ensemble classifier was developed utilizing a soft voting scheme. This ensemble comprised three distinct classifiers: Random Forest, Logistic Regression, and a Multilayer Perceptron (MLP). The inclusion of these classifiers was strategically motivated by their complementary strengths: Random Forest's capacity to model nonlinear interactions and handle noisy data; Logistic Regression's interpretability and effectiveness in modeling linear relationships; and MLP's ability to learn complex, high-order feature interactions. This design reflects ensemble learning theory, which posits that diversity among base learners enhances generalization and mitigates individual model biases (Zhou, 2012).

The efficacy of the hybrid ensemble was empirically validated. It achieved an accuracy of 96.1% and an F1-score of 96.0%, outperforming the best standalone model (Random Forest with all features) both in terms of accuracy and computational efficiency. The ensemble’s training time was reduced by approximately 30%, demonstrating the scalability of the method. These findings reinforce the theoretical expectations from the literature, where hybrid and heterogeneous ensemble models have been shown to outperform single-model approaches in terms of accuracy, robustness, and resilience to overfitting (Wang, 2012; Hu & Maybank, 2008).

Moreover, the modularity of the proposed method renders it particularly suitable for deployment in dynamic, real-world e-commerce systems. Logistic Regression's explainability supports business transparency, while the ensemble's speed and accuracy facilitate timely decision-making in recommendation engines and personalized marketing platforms. This is especially pertinent in the e-commerce domain, where the ability to act upon predictive insights in real time can directly impact conversion rates and revenue generation (Komarek, 2004).

In summary, this hybrid method demonstrates that combining targeted feature selection with a diverse ensemble of classifiers yields a model that is not only more accurate but also more efficient and interpretable. It exemplifies the practical potential of integrating theoretical advancements in ensemble learning with application-driven design principles in data mining.

References

Amin, M. Z., & Ali, A. (2017). Application of Multilayer Perceptron (MLP) for Data Mining in Healthcare Operations. *3rd International Conference on Biotechnology*, Lahore, Pakistan.

Bhargavi, P., & Jyothi, S. (2009). Applying Naive Bayes Data Mining Technique for Classification of Agricultural Land Soils. *IJCSNS International Journal of Computer Science and Network Security*, 9(8), 117–122.

Gupta, B., Rawat, A., Jain, A., Arora, A., & Dhami, N. (2017). Analysis of Various Decision Tree Algorithms for Classification in Data Mining. *International Journal of Computer Applications*, 163(8), 15–20.

Heckerman, D. (1996). Bayesian networks for data mining. *Data Mining and Knowledge Discovery, 1*(1), 79–119.

Hu, W., Hu, W., & Maybank, S. (2008). AdaBoost-Based Algorithm for Network Intrusion Detection. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 38(2), 577–582.

Kataria, A., & Singh, M. D. (2013). A Review of Data Classification Using K-Nearest Neighbour Algorithm. *International Journal of Emerging Technology and Advanced Engineering*, 3(6), 354–359.

Komarek, P. (2004). Logistic Regression for Data Mining and High-Dimensional Classification. *Carnegie Mellon University*.

Peling, I. B. A., Arnawan, I. N., Arthawan, I. P. A., & Janardana, I. G. N. (2017). Implementation of Data Mining to Predict Period of Students Study Using Naive Bayes Algorithm. *International Journal of Engineering and Emerging Technology, 2*(1), 53–55.

Priyam, A., Abhijeet, A., Gupta, R., Rathee, A., & Srivastava, S. (2013). Comparative Analysis of Decision Tree Classification Algorithms. *International Journal of Current Engineering and Technology, 3*(2), 334–336.

Sharma, H., & Kumar, S. (2016). A Survey on Decision Tree Algorithms of Classification in Data Mining. *International Journal of Science and Research, 5*(4), 2094–2100.

Vembandasamy, K., Sasipriya, R. R., & Deepa, E. (2015). Heart Diseases Detection Using Naive Bayes Algorithm. *International Journal of Innovative Science, Engineering & Technology, 2*(9), 441–447.

Wang, R. (2012). AdaBoost for Feature Selection, Classification and Its Relation with SVM: A Review. *Physics Procedia*, 25, 800–807.

Zhou, Z.-H. (2012). Ensemble Methods: Foundations and Algorithms. Chapman and Hall/CRC.