Vella (2023)

## Summary:

The abstract showcases the study’s goal of assessing the impact of local education within the ICT sector both vocational and traditional meaning MCAST and UOM respectively. It emphasizes the skill gaps that exist between the educational backgrounds and industry needs. The study seeks to bridge the gap by collecting insights with ICT professionals, educators, and industry stakeholders. The researcher uses a mixed method approach that includes data scraping 200 LinkedIn profiles paired with qualitative analysis specifically interviews and surveys using the software MAXQDA.

## Literature:

This section outlines relevant studies pertaining ICT education, skill gap and industry requirements. It emphasises the importance of industry-academia cooperation and stresses the urgent need for practical skills.

## Methodology:

Vella uses a mixed method approach more specifically a triangulation of research methods. A mix of qualitative and quantitative data was used.

Data collection:

* LinkedIn Scraping: 200 ICT professionals were chosen and their LinkedIn profiles were scraped, 100 of them possess a vocational educational background, MCAST and 100 of them possess a traditional educational background UOM.
* To guarantee quality and applicability, the scraped data was thoroughly cleaned.
* After the data was cleaned, MAXQDA was used for qualitative analysis, profiles were coded using themes such as career pathways and skill set. A thematic comparison using a code matrix and case models were used to organise the data by educational institution MCAST vs UOM.
* To combine qualitative and quantitative insights, FAMD (Factorial Analysis for Mixed Data) was used to highlight the disparities in career paths and skills sets between traditional and vocational graduates.
* Interviews with 7 ICT professionals were conducted and later analysed using MAXQDA once again.
* Two surveys were created, one for students undergoing an ICT qualification and a second online survey was aimed at ICT graduates who are either currently working in the field or have previous experience in the ICT industry. Amount of respondents was not specified.

## Pros:

* Research Method: Triangulation of methods is useful for my research as I too will be using qualitative, quantitative and a prototype.

Tamhankar et al(2018)

## Abstract:

In order to enhance assortment planning for a shop handling products with low demand, the study looks into clustering and predictive modelling techniques. The study contrasts several clustering techniques and investigates how these clusters affect the effectiveness of predictive models of the regression and classification types. Three main topics will be addressed by the study: how well clustering creates valuable product segments, how clustering affects predicted accuracy, and whether situations require the employment of regression or classification techniques.

## Literature:

The researchers examine a number of assortment planning techniques, such as conventional models based on demand elasticity, space elasticity, and substitution. By enhancing demand estimates, they demonstrate how machine learning algorithms in particular, clustering when paired with predictive analytics can aid in decision-making. According to the study, traditional methods have trouble predicting with precision when there is little demand, particularly for businesses who have large and varied product assortments.

## Methodology:

Clustering models:

* K-Means: This well-known clustering algorithm was used to divide products into groups according to characteristics such as lifecycle stages and failure rates. An elbow plot (figure 2 in paper) was used to find clusters that are homogeneous in demand patterns and product lifecycles, hence determining the ideal number of clusters.
* HDBSCAN: A density-based clustering method that excludes sparse points as noise and groups data points with lots of neighbours. Although HDBSCAN produced clusters, it limited the dataset's interpretability by classifying a significant amount of it as noise.
* Kohonen (Self-organising Maps): A clustering method based on neural networks that groups products according to their failure rates and lifecycles while reducing the complexity of the data. Nevertheless, it added minimal fresh data because its clusters mostly overlapped with K-Means'.

K-Means was chosen as the final clustering technique based on the findings.

Predictive Models:

* Classification Models: Logistic regression, classification trees (CART), linear discriminant analysis (LDA), and more sophisticated models like bagged trees, boosted logistic regression, and multilayer perceptron (MLP) neural networks are examples of classification models. Bagged Logistic Regression was the best-performing classification model (with an Area Under Curve (AUC) score of 0.9841 in the absence of clustering.)
* Regression Models: To precisely forecast the quantity of units sold, the study also created regression models. Neural networks, Zero-Inflated Poisson Regression (ZIP), and Multiple Linear Regression (MLR) were employed. With a high adjusted R-squared (90.5%), ZIP Regression was the best at forecasting sparse demand, which made it perfect for situations where there were a lot of zero sales.

## Pros:

* Effective use of Clustering: K-Means created interpretable segments.
* K-means could be a potential algorithm used in this project

## Cons:

* Findings are based on data from a single retailer.
* Limitations of HBDSCAN and Kohonen Networks, both methods struggled when handling sparse data, either overlapping with the K-means results or classifying too many data points.

Shaikh et al (2021)

## Abstract:

This study examines how big data can be used to optimise assortment strategies in Pakistan's organised retail industry. It evaluates the contributions of large data, sophisticated algorithms, and knowledgeable data scientists to assortment planning. Using survey data from IT professionals in the retail industry and Partial Least Squares Structural Equation Modelling (PLS-SEM), the study concludes that although big data is important, sophisticated algorithms are necessary to fully realise its potential. These results are considerably impacted by skilled data scientists, especially in situations that call for quick and intricate analysis.

## Literature:

The literature study emphasises how big data has the ability to revolutionise the retail industry by influencing strategic choices like inventory and layout optimisation through information on consumer behaviour and preferences. Important elements consist of:

* Big Data in Retail: By examining trends like purchase quantity, frequency, and outlet location, big data helps with real-time decision-making, particularly in inventory management and assortment planning.
* Advanced Algorithms: Although there are still difficulties because of the complexity of the data and the requirement for qualified analysts to properly understand the results, algorithms are essential for gathering useful insights from complicated data.
* Role of Date Scientists: The proficiency of data scientists is also essential for the efficient use of big data, especially when it comes to creating actionable insights in dynamic retail settings.

## Methodology:

The study uses a quantitative methodology, testing hypotheses with PLS-SEM:

* Sampling Methods: Quota sampling was used to gather data from 100 IT professionals employed in Pakistan's organised retail FMCG industry. This choice supports the theory-building goal of the investigation.
* Data Analysis: Since smartPLS is appropriate for small, non-normally distributed samples, it was used to do structural equation modelling (SEM).
* R-Squared Values: assortment techniques had an R-squared of 0.649, showing a minor relationship, while advanced algorithms had an R-squared of 0.704, indicating a large effect.

## Pros:

* Particular Attention to Retail Assortment, deepens knowledge of assortment planning, a topic that has a big influence on retail profitability.

## Cons:

* Dependency on Quota Sampling: The study may be biased because it employs quota sampling as opposed to probability sampling. This approach would reduce the representativeness of the sample, which would compromise the validity of the findings, particularly when the industry sees new trends in the adoption of technology.
* Technology and Resource Limitations: A large number of Pakistani businesses lack the infrastructure and resources necessary to fully deploy sophisticated big-data analytics. This restriction raises the possibility that the results may represent a perfect situation that isn't always possible for all local merchants.
* Geographic context: The findings' in this study compared to other cultural and economic contexts is limited by the focus on Pakistan's retail sector. Countries can differ greatly in terms of retail dynamics, consumer preferences, and data management infrastructure, which could affect how well big data and algorithmic approaches work.

Kurniawen et al (2018)

## Abstract:

By using transaction data from a supermarket, Kurniawan et al. (2018) investigate how market basket analysis (MBA) might reveal consumer buying patterns. The Apriori algorithm was used by the researchers to create a desktop application that generates association rules. 30 rules with confidence and support values of 46.69% and 1.78%, respectively, were produced by testing on a dataset from an Indonesian supermarket. By providing insights into item associations and buying trends, the study illustrates how MBA may be used to optimise stock management and product assortment.

## Literature:

The researchers discuss the importance of comprehending consumer behaviour, particularly through transaction data, which provides information about preferences, item associations, and buying patterns. Among the main topics discussed are:

* Market Basket Analysis: Known as a data mining technique, MBA helps firms identify which products are frequently purchased together, enabling them to develop well-informed assortment and promotional tactics.
* Association Rules: According to the study, retailers can use association rules such as "If item A is bought, item B is also likely to be bought" to organise product layouts, make restocking choices, and run targeted promotions.
* Understanding consumer behaviour and creating strategies that adapt to the ever changing requirements and preferences of consumers are made easier with the help of this study.

## Methodology:

Kurniawan et al. use the Apriori algorithm for mining association rules, which is a popular option for MBA because of its effectiveness in locating frequently occurring item sets. Important elements consist of:

* Data collection: Transaction data from a particular time period, including receipts from more than 1,553 transactions from a supermarket in Malang, Indonesia, is used in the study.
* Application Testing: 56 sample transactions with 20 items each were used to test the constructed application. These transactions were utilised to generate 30 association rules that represent purchase patterns and to calculate support and confidence levels.

## Pros:

* Practical Application: By illustrating how merchants can optimise assortments based on real purchase trends, the study shows how MBA is applicable to real-world retail situations.
* Use of Apriori Algorithm is another potential algorithm that can be used for this study

Kaur et al (2016)

## Abstract:

In order to determine consumer buying patterns, Kaur and Kang (2016) conducted a study on Market Basket Analysis (MBA) utilising association rule mining. Its goal is to assist retailers in better understanding consumer behaviour and decision-making. The authors suggest a novel method that dynamically takes into account variations in transaction data over time, in contrast to static algorithms. By detecting both high-confidence rules and outliers, this strategy enables merchants to maximise revenues and adjust to changing client preferences.

## Literature:

The study offers a thorough rundown of data mining methods, with a focus on retail environments:

* An important data mining approach used in MBA programs is association rule mining, which is used to identify connections between products that are frequently purchased together. Apriori and FP-Growth are two well-known algorithms whose association rules are represented by metrics like confidence and support.
* Applications of MBA: The study highlights how MBAs are used in a variety of industries, but particularly in retail, where they aid in customer relationship management, inventory control, and cross-selling.
* Outlier Detection: Unusual buying patterns may be represented as outliers in transaction data. Retailers can spot irregularities that could point to changing patterns or even fraud by identifying these outliers.

## Methodology:

The paper's suggested algorithm combines change modelling and association rule mining:

* Data Set: For their tests, the authors employed an extensive bakery dataset that was split up into four time frames, each of which had 2,000 transactions.
* Algorithm-ARM-Predictor: This algorithm captures association rules in successive time frames by building upon the Apriori algorithm.
* ARM-Predictor:
  + Maintains a score table that records changes in high-confidence associations throughout time.
  + Distinguishes between rules that are likely to become obsolete and those that have a high degree of predictability for upcoming transactions.
  + Rules from every new time frame are periodically added to the Score Table, which is then updated to reflect changing patterns and assign scores based on confidence criteria.
* Outlier Detection: The algorithm finds outliers, or rules that no longer satisfy the confidence level, after determining association scores. Retailers can use this to identify trends that may be fading in importance.

## Pros:

* Effective Outlier identification: By incorporating outlier identification, retailers can find products with diminishing relationships, which is helpful for inventory management.

Kanagaraj et al (2016)

## Abstract:

The application of modern data analytics to enhance business performance in speciality retail sectors is examined in this study. In order to find high-performing items, comprehend client purchase patterns, and guide strategic decision-making, the authors stress the significance of sales data analysis. The study shows how interactive data visualisation can give firms insights into sales patterns, geographic dispersion, and market performance using a Power BI dashboard, SQL queries, and the DAX query language.

## Literature:

Kanagaraj et al. investigate a number of data analysis methods and techniques:

* Data Preparation: In order to guarantee clean and trustworthy data, the study emphasises the importance of data preprocessing (such as deduplication and standardisation) utilising programs like Microsoft SQL Server Integration Services (SSIS).
* Exploratory Data Analysis (EDA): EDA enables researchers to find important patterns in sales data, like customer segmentation and seasonal trends, by using visualisation libraries like Seaborn.
* Advanced Analytics and DAX: Creating dynamic dashboards that display real-time data insights requires the use of Power BI's DAX (Data Analysis Expressions) query language, which makes custom computations and aggregations easier.
* Data Integration: With applications for resource allocation and inventory management, the study explores merging SQL and DAX within Power BI to offer a strong framework for assessing business performance.

## Methodology:

In order to analyse and visualise sales data, the process entails building a Power BI dashboard that integrates SQL and DAX:

* Data Collection: With a focus on historical and geographic data dimensions, sales and distribution data were gathered from speciality retail establishments.
* Data Preparation: To address missing values, duplication, and format inconsistencies, the data was cleaned, standardised, and pre-processed using SQL and SSIS.
* Dashboard Development:
  + DAX Query Language: Metrics like profit, net sales, and gross sales were computed using custom DAX methods. Interactive visualisations that show sales by time period, area, and product category were made possible by the enquiries.
  + SQL Enquiries to Verify Data: Before being loaded into Power BI, the data was verified and filtered using SQL queries, guaranteeing its integrity and accuracy.

## Pros:

* Real-Time Visualisations: In order to monitor continuing trends and make necessary strategy adjustments, Power BI's interactive graphics provide real-time updates. May be a potential research method to showcase results or data.

Santos (2019)

## Abstract:

A real-world assortment optimisation issue for an online grocery store is tackled by Santos (2019). The study focusses on how to reduce assortment size without significantly lowering sales in order to choose the ideal number of unique Stock Keeping Units (SKUs) for a certain product subcategory, in this case rice. The objective is to reduce stock outs and increase revenue. The suggested approach uses transactional log data to estimate customer behaviour and revenue potential for each assortment, taking into account variables like SKU price, demand, stock out rates, and customer preferences.

## Literature:

Key themes in assortment optimisation are examined in the literature review, with particular attention to:

* Consumer Perception of Assortment: Research indicates that customer satisfaction is influenced by how varied an assortment is perceived to be. When properly handled, a decrease in assortment can make decisions easier and possibly increase sales.
* Choice Models: Various parametric and non-parametric choice models are reviewed in this paper. Commonly used multinomial logit (MNL) and nested logit models may have over-fitting or under-fitting problems, particularly when substitution effects are included.
* Methods of Optimisation: The study looks at techniques like Mixed-Integer Optimisation (MIO) and Mixed-Integer Linear Programming (MILP) as workable answers for assortment optimisation that is revenue-focused. Santos's approach is based on Bertsimas and Mišic's (2015) MIO model, which uses ranking-based consumer preferences.

## Methodology:

The approach modifies Bertsimas and Mišic's MIO model to meet particular business requirements:

* Data Collection: During the first half of 2018, transactional data from the retailer was examined with an emphasis on the rice subcategory.
* Factors and Limitations:
  + Calculating Revenue: Both SKU pricing and stockout level are taken into account by the model. As a penalty factor, stockouts lower the potential income for SKUs with a high stockout rate.
  + Brand Constraint: Each brand must have at least one SKU in the assortment to preserve brand diversity.
  + Model of Optimisation: With binary decision variables that indicate whether an SKU should be kept (1) or taken (0) out of the assortment, the model is configured as a MILP. Transactional data is used to determine customer preferences, and each SKU is ranked according to sales volume and frequency.

Karki (2018)

## Abstract:

In order to manage retail selection, Karki's study suggests a hybrid model that incorporates consumer behaviour into product classification. Understanding that some items have a greater influence on consumer loyalty and buying habits than others, the study optimises product assortment by combining association rule mining and clustering. In particular, it uses Weighted Association Rule Mining (WARM) to create an assortment strategy that ranks products according to their impact on consumer behaviour, Fuzzy C-means clustering for product classification, and K-means clustering for customer segmentation.

## Literature:

The literature analysis highlights the drawbacks of conventional assortment management, which frequently prioritises sales-based KPIs over customer behaviour. Among the main themes are:

* The Role of Customer Behaviour in Assortment Planning: Research indicates that retention can be enhanced by comprehending how product diversity and customer preferences interact.
* Customer-Product Relationships: Studies by Hebblethwaite et al. (2017) and Lariviere and Van den Poel (2004) demonstrate how particular items can increase client loyalty and lower attrition.
* Segmentation and Clustering: Previous research on customer segmentation using k-means clustering and RFM (Recency, Frequency, Monetary) models supports the notion that arranging clients according to value improves the targeted assortment.

## Methodology:

The study contains multiple phases and adheres to the CRISP-DM framework:

* Gathering and Preparing Data:
  + More than 26,000 distinct transactions are included in the dataset, which was obtained from UCI's Machine Learning Repository.
  + With qualities pertinent to consumer behaviour and product properties, the data was cleaned and organised into a "master" database for analysis.
* K-Means Clustering for Customer Segmentation:
  + RFM measurements were used to segment the customer base in order to find high-value groups. Three primary client segments: Star, Loyal, and Inactive customers were identified with the aid of K-means clustering.
  + The best number of clusters was found using the elbow approach, and the most engaged consumers were represented by the top segments (Loyal and Star).
* Classifying Products Using Fuzzy C-Means Clustering:
  + Six groups were created by grouping products according to factors including how often they were purchased. Soft clustering, which reflects real-world overlaps in customer preferences, was made possible by fuzzy C-means.
  + Products could be a part of numerous clusters with different levels of membership.
* Weighted ARM and Association Rule Mining (ARM):
  + Frequent itemsets and association rules between customer segments and product clusters were found by traditional ARM utilising the Apriori algorithm.
  + Weighted ARM (WARM) helped guarantee that important products with lower sales still receive precedence in the assortment by introducing minimal support levels and weights to highlight products that affect consumer loyalty.

## Pros:

* Dataset is public and available to use on UCI (Chen et al 2012)( <https://archive.ics.uci.edu/dataset/352/online+retail>)

|  |  |  |
| --- | --- | --- |
| Paper | Algorithms/ Solutions used | Purpose |
| Vella (2023) | LinkedIn scraping, MAXQDA, Factorial Analysis for Mixed Data (FAMD) | Analysed education’s impact on ICT professionals, using LinkedIn data for quantitative insights and MAXQDA for a thematic analysis. |
| Tamhankar et al.(2018) | K-means Clustering,HDBSCAN, Kohonen Networks, Bagged Logistic Regression, Zero-inflated Poisson regression(ZIP) | Applied clustering and predictive analytics to optimise product assortments for demand forecasting. |
| Shaikh et al.(2021) | Partial Least Squares Structural Equation Modeling (PLS-SEM), Big Data algorithms, Advanced Predictive models | Analysed the role of big data and advanced algorithms in optimising product assortment and inventory management. |
| Kurniawan et al. (2018) | Apriori Algorithm (Association Rule Mining) | Utilised market basket analysis to uncover frequent item sets, aiding in customer behaviour. |
| Kaur and Kang (2016) | ARM-Predictor Algorithm (dynamic association rule mining) | Developed a time-adaptive approach to track changing market trends and evolving customer preferences. |
| Kanagaraj (2023) | SQL, Power BI, DAX (Data Analysis Expressions), Exploratory Data Analysis (EDA) with Seaborn | Built a data visualisation dashboard for real-time insights into sales, inventory, supporting strategic decision-making. |
| Santos (2019) | Mixed-Integer Linear programming (MILP), Bertsimas and Misics’s MIO model | Applied MILP for SKU optimisation in online retail, reducing assortment size while retaining sales potential. |
| Karki (2018) | K-means Clustering, Fuzzy C-means Clustering, Apriori Algorithm (ARM), Weighted Association Rule Mining (WARM) | Create a hybrid model combining customer segmentation and product clustering to tailor assortment strategies to behaviour. |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Author of Paper | Number of observations | Number of variables | Is Public? | Release Date |
| Vella | 137 | 12 | No but sourced from public data | 2023 |
|  | 100 |  | No (personal survey) |  |
|  | 6 |  | No(personal interview) |  |
| Tamhankar | Over a two-year time frame | 3 + not disclosed | No | 2018 |
| Shaikh | 100 | Not Disclosed | No | / |
| Kurniawan | 1553 | 5 | No | 2018 |
| Kaur | 2000 | / | Yes (but link does not open) | / |
| Santos | / | 3 | No | / |
| Karki /Chen | 540,000 | 6 | Yes | 2015 |
| Safara | / | 11 | No but sourced from public data | 2022 |
| Rusniati | 100 | 3 | No (Personal Survey) | 2024 |
| Pyzalska | 26,568 | / | No | 2020 |
| Khatter | * 1st dataset= / * 2nd dataset=2 million + | * 8   4 | * Yes   Yes | * 2016   2018 |
| Huang | / | * 3 | * No | * 2014 |
| Farkas | / | / | No | / |
| Desara | 222 | 7 | No (Personal Survey) | 2021 |

