INTERNATIONAL UNIVERSITY OF EAST AFRICA
'Learning to Succeed'

FACULTY OF SCIENCE AND TECHNOLOGY

END OF SEMESTER EXAMINATIONS - DECEMBER 2025

PROGRAMME: MIT
YEAR/SEM: YEAR 2/SEMESTER 1
COURSE CODE: MIT812
NAME: BIG DATA ANALYTICS

DATE:
TIME:

INSTRUCTIONS TO CANDIDATES:

SECTION A

Attempt One Question from Section A

Question 1:

QUESTION ONE (40 MARKS)

i. Define Big Data and explain why traditional data processing tools struggle to handle it. (4 marks)

ii. Briefly describe the ‘Volume’ and ‘Variety’ characteristics of Big Data, giving one example for each. (4 marks)

iii. Briefly describe the ‘Velocity’ and ‘Veracity’ characteristics of Big Data, giving one example for each. (4 marks)

iv. Explain the ‘Value’ dimension of Big Data and provide one example from either business or the public sector. (4 marks)

v. Provide two real-world scenarios (from any sectors) showing how at least three of the 5Vs appear together in each scenario. (4 marks)

vi. Compare traditional RDBMS-based data processing with distributed Big Data architectures in terms of scalability and fault tolerance. (4 marks)

vii. Compare traditional and distributed architectures in terms of storage architecture and the types of workloads they handle best. (4 marks)

viii. Explain what Hadoop is and describe the roles of HDFS and MapReduce in its ecosystem. (4 marks)

ix. Explain what Apache Spark is and describe the roles of RDDs/DataFrames and in-memory processing. (4 marks)

x. Give one appropriate use case for Hadoop as a batch processing framework and justify why Hadoop is suitable for that case. (4 marks)

Question 2:

QUESTION TWO (40 MARKS)

i. Give one appropriate use case for Hadoop as a long-term storage/archival platform and justify your answer. (4 marks)

ii. Provide two distinct use cases for Spark (for example, streaming analytics or iterative machine learning) and justify why Spark is suitable in each case. (4 marks)

iii. Distinguish between supervised and unsupervised learning, giving one practical example of each. (4 marks)

iv. Explain the concept of semi-supervised learning and describe a situation where it is more suitable than purely supervised or unsupervised learning. (4 marks)

v. Choose one application area (e.g. credit scoring, image recognition or disease prediction) and explain why it is best modelled as a supervised learning problem. (4 marks)

vi. Summarise the main stages of a Big Data analytics pipeline (data acquisition, cleaning, transformation, modelling, evaluation, deployment), giving a brief note on each.(4 marks)

vii. Discuss two key challenges that may be encountered during data acquisition and data cleaning in a Big Data project, and suggest strategies to address them. (4 marks)

viii. Discuss two key challenges that may occur during modelling, evaluation, or deployment stages in a Big Data project, and suggest strategies to address them. (4 marks)

ix. Define a Data Lake and a Data Warehouse, and explain the difference between ‘schema-on-read’ and ‘schema-on-write’ with examples. (4 marks)

x. With suitable examples, differentiate Batch Processing from Stream Processing and name one real-world application where each approach is most appropriate. (4 marks)

SECTION B

Attempt any One Question from Section B

Question 1:

QUESTION ONE (60 MARKS)

Uganda faces persistent youth unemployment, skills mismatch, and the absence of a comprehensive national talent management system. To address this, the Ministry of Gender, Labour and Social Development has developed the National Talent and Employment Analytics Platform (N-TEAP). The platform integrates data from job seekers, employers, URA, NSSF, universities, migration records, and labour market surveys. You have been engaged as a Big Data Analyst to use a dataset of 10,000 youth records extracted from N-TEAP to generate insights that can inform national policy and guide system enhancements.

You can also obtain a relevant youth unemployment or labour market analytics dataset from reputable online open-data platforms such as Kaggle, the World Bank Open Data Portal, ILOSTAT, or any other credible source. Ensure the dataset has sufficient variables to allow descriptive analytics, inferential statistics, predictive modelling, clustering, and visualisation in R.

TASKS:

i. Exploratory and Descriptive Analysis (15 marks)

Using appropriate numerical summaries and visual displays, explore and describe key patterns in the dataset related to employment status, time taken to secure employment, education, skills, certifications, region, gender, and sector of interest. Present notable trends and anomalies.

ii. Relationship and Group Comparison Analysis (10 marks)

Investigate differences in employment outcomes across groups (gender, education, region, training participation). Use suitable statistical techniques to assess differences and quantify key relationships. Justify your method choices.

iii. Predictive Modelling of Employment Outcomes (15 marks)

Develop and evaluate predictive models to determine factors influencing employment likelihood and job search outcomes (time to employment, interviews, income). Report model performance and practical implications.

iv. Segmentation and Profiling of Youth (Clustering) (8 marks)

Use unsupervised learning to segment youth based on skills, education, experience, and job search behaviour. Describe segments and how each could be targeted in policy or training interventions.

v. Data Visualisation for Decision Support (6 marks)

Design visualisations suitable for embedding into an N-TEAP dashboard. Provide examples generated in R and justify their decision-support relevance.

vi. Policy and System Recommendations (6 marks)

Provide evidence-based recommendations on reducing unemployment and improving N-TEAP functionalities. Link recommendations to empirical findings.

Question 2:

QUESTION TWO (60 MARKS)

Uganda's education sector continues to struggle with challenges affecting access, quality, equity, and learning outcomes. Major issues include high dropout rates, poor literacy and numeracy performance, teacher absenteeism, overcrowded classrooms, gender disparities, weak monitoring systems, and low transition rates. To support data-driven reforms, the Ministry of Education has developed the Education Data Intelligence Platform (EDIP), integrating EMIS, UNEB, teacher attendance, census, inspection, e-learning, and financial data. You have been appointed as a Big Data Analyst to investigate performance patterns, system inefficiencies, and learning disparities using a dataset of at least 10,000 records spanning students, teachers, and schools. You can also to obtain a relevant education-sector dataset from trusted open-data sources such as Kaggle, the World Bank Open Data Portal, UNESCO Institute for Statistics, or any other credible repository. The dataset must contain adequate variables to support descriptive analytics, inferential statistics, predictive modelling, machine learning, clustering, and data visualisation in R.

TASKS:

i. DESCRIPTIVE ANALYTICS (12 Marks)

Conduct an exploratory analysis focusing on key education indicators such as student performance, attendance, resource allocation, school-level conditions, and demographic characteristics. Identify notable disparities across districts, regions, or school types.

ii. INFERENTIAL ANALYTICS (18 Marks)

Examine relationships and differences among variables related to learning outcomes, school characteristics, and teacher-related factors. Use suitable statistical techniques to test relationships, compare groups, and draw evidence-based conclusions.

iii. PREDICTIVE MODELLING (12 Marks)

Develop predictive models to estimate key education outcomes (e.g., student performance, dropout patterns, or school efficiency). Evaluate model accuracy and interpret the importance of predictors in practical, policy-relevant terms.

iv. MACHINE LEARNING & CLUSTERING (8 Marks)

Apply appropriate machine learning or clustering techniques to uncover hidden patterns or segment schools, learners, or teachers into meaningful groups that can guide targeted interventions.

v. DATA VISUALISATION (6 Marks)

Generate clear, decision-oriented visualisations in R that communicate critical insights to policymakers. Choose suitable plotting methods to highlight disparities, trends, and risk patterns.

vi. POLICY & SYSTEM RECOMMENDATIONS (4 Marks)

Based on empirical findings, propose actionable policy recommendations and suggest potential enhancements for EDIP to improve monitoring, quality assurance, and data-driven decision-making.

Question 3:

QUESTION THREE (60 MARKS)

Uganda's Ministry of Health faces persistent health-system challenges including disease outbreaks, maternal and child health concerns, medicine stockouts, inefficient referral pathways, overcrowded facilities, rising non-communicable diseases, and limited integration of health information systems. To address these issues, the Ministry has developed the National Health Intelligence and Analytics Platform (NHIAP), integrating DHIS2, EMR systems, laboratory information systems, LMIS, VHT mobile data, NIRA, NHIS claims, ambulance GPS logs, environmental datasets, and UBOS demographic records. You have been contracted as a Big Data Analyst to examine service delivery performance, disease trends, and patient outcomes using a dataset of at least 10,000 health records. You can also obtain a relevant health-sector dataset from reputable open-data platforms such as Kaggle, WHO Global Health Observatory, World Bank Open Data, or any other credible repository. The dataset must contain adequate variables to support descriptive analytics, inferential statistics, predictive modelling, machine learning, clustering, and data visualisation using R.

TASKS:

i. DESCRIPTIVE ANALYTICS (12 Marks)

Conduct an exploratory descriptive assessment of key health indicators such as patient demographics, disease distribution, service utilisation, facility performance, and treatment outcomes. Highlight notable disparities across districts, facility levels, or population groups.

ii. INFERENTIAL ANALYTICS (18 Marks)

Investigate statistical relationships and differences among selected variables related to disease categories, patient outcomes, facility characteristics, and system performance. Apply suitable inferential methods to draw evidence-based conclusions.

iii. PREDICTIVE MODELLING (12 Marks)

Develop predictive models to estimate key health outcomes such as length of stay, patient risk levels, ambulance response patterns, or mortality likelihood. Evaluate model accuracy and explain the significance of identified predictors.

iv. MACHINE LEARNING & CLUSTERING (8 Marks)

Apply appropriate machine learning and clustering approaches to identify facility performance groupings, disease pattern clusters, patient risk categories, or service delivery profiles that can guide policy and resource allocation.

v. DATA VISUALISATION (6 Marks)

Produce clear and well-structured R visualisations capable of supporting health policy decisions. Use appropriate graphical techniques to communicate trends, disparities, or system inefficiencies.

vi. POLICY & SYSTEM RECOMMENDATIONS (4 Marks)

Using insights from your analysis, provide well-justified policy recommendations and propose actionable improvements to NHIAP to strengthen data-driven health system performance.

Question 4:

QUESTION FOUR: (60 MARKS)

Uganda continues to face persistent poverty and rising inequality despite programmes such as PDM, YLP, UWEP, Emyooga, and SAGE. Key challenges include rural poverty, youth unemployment, limited financial inclusion, weak monitoring of government livelihood programmes, climate-related shocks, and inadequate visibility into household welfare dynamics. To strengthen evidence-based planning, the Ministry of Finance, Planning and Economic Development developed the National Socio-Economic Analytics Platform (N-SEAP), integrating household survey data, URA tax records, mobile money transactions, agricultural statistics, satellite climate data, NIRA records, digital financial usage logs, and geospatial community data.You have been hired as a Big Data Analyst to examine poverty dynamics, vulnerability patterns, and welfare disparities using a dataset of at least 10,000 household- and community-level records. You can obtain a relevant socio-economic or poverty-related dataset from reputable open-data platforms such as Kaggle, World Bank Open Data, UNDP Human Development Data, FAO datasets, or any other credible repository. The dataset must contain adequate variables to support descriptive analytics, inferential statistics, predictive modelling, machine learning, clustering, and advanced data visualisation in R.

TASKS:

i. DESCRIPTIVE ANALYTICS (12 Marks)

Perform descriptive and exploratory analytics on key socio-economic indicators such as household welfare, income, expenditure, financial inclusion, vulnerability indices, climate shock exposure, and demographic characteristics. Identify notable disparities across regions, household types, or livelihood categories.

ii. INFERENTIAL ANALYTICS (18 Marks)

Using suitable inferential approaches, investigate relationships and differences among variables related to income distribution, poverty likelihood, demographic characteristics, climate vulnerability, or financial inclusion. Draw statistically defensible conclusions that support policy interpretation.

iii. PREDICTIVE MODELLING (12 Marks)

Develop predictive models capable of estimating key outcomes such as welfare status, expenditure levels, vulnerability risk, or poverty probability. Evaluate model performance and interpret the role of major predictors in the context of socio-economic planning.

iv. MACHINE LEARNING & CLUSTERING (8 Marks)

Apply appropriate machine learning or clustering techniques to identify distinct groups of households or communities based on welfare conditions, vulnerability patterns, financial behaviour, climate risk profiles, or livelihood sources. Explain the policy relevance of the discovered segments.

v. DATA VISUALISATION (6 Marks)

Produce clear, meaningful, and decision-oriented visualisations in R that effectively communicate trends, patterns, disparities, or emerging risks within the socio-economic data. Ensure that visual choices match the analytical objective.

vi. POLICY & SYSTEM RECOMMENDATIONS (4 Marks)

Based on empirical findings, provide actionable policy recommendations and propose enhancements for N-SEAP to support targeted poverty reduction, improved monitoring, and data-driven socio-economic management.

Question 5:

QUESTION FIVE (60 MARKS)

Uganda's agricultural sector continues to experience major structural challenges, including low crop yields, high postharvest losses, climate-related disruptions, weak market access, limited mechanisation, poor disease surveillance, and fragmented value chains. To strengthen evidence-based agricultural planning, the Ministry of Agriculture, Animal Industry and Fisheries (MAAIF) developed the Agricultural Data and Intelligence Platform (Agri-DIP), integrating satellite imagery, climate data, production statistics, UBOS surveys, livestock disease reports, market price feeds, mobile-money activity, extension service logs, and digital agritech platform data. You have been contracted as a Big Data Analyst to examine agricultural productivity, climate vulnerability, market efficiency, and postharvest loss patterns using a dataset of at least 10,000 agricultural production and market records. You can obtain a relevant agriculture-related dataset from trusted open-data platforms such as Kaggle, FAO datasets, World Bank Open Data, GEOGLAM agricultural datasets, or any other credible repository. The dataset selected must contain adequate variables to support descriptive analytics, inferential analysis, predictive modelling, machine learning, clustering, and data visualisation in R.

TASKS:

i. DESCRIPTIVE ANALYTICS (12 Marks)

Conduct descriptive and exploratory analytics on key agricultural indicators such as crop yields, weather conditions, production inputs, market dynamics, postharvest losses, disease prevalence, and farmer characteristics. Highlight important disparities across regions, crops, farmer types, or climatic zones.

ii. INFERENTIAL ANALYTICS (18 Marks)

Apply suitable inferential techniques to investigate relationships, group differences, and patterns involving yields, climate variables, market performance, postharvest behaviour, and production characteristics. Provide statistically meaningful interpretations that support agricultural decision-making.

iii. PREDICTIVE MODELLING (12 Marks)

Develop predictive models for outcomes such as crop yield, farmer income, loss probability, disease risk, or market price fluctuations. Evaluate model performance and interpret the significance of key predictors in the context of agricultural productivity and resilience.

iv. MACHINE LEARNING & CLUSTERING (8 Marks)

Use appropriate machine learning or clustering approaches to identify meaningful agricultural segments such as farmer groups, climatic zones, market clusters, or farm-level risk profiles. Explain the policy and operational relevance of the derived clusters.

v. DATA VISUALISATION (6 Marks)

Develop clear, policy-ready visualisations in R that communicate trends, disparities, anomalies, and emerging risks within the agricultural ecosystem. Justify each visualisation technique in relation to the insight it communicates.

vi. POLICY & SYSTEM RECOMMENDATIONS (4 Marks)

Based on empirical findings, provide actionable recommendations for MAAIF. Suggest improvements for Agri-DIP to support real-time analytics, better monitoring of postharvest losses, enhanced climate-risk planning, and data-driven value-chain modernisation.

Question 6:

QUESTION SIX (60 MARKS)

Uganda is experiencing a rapidly rising burden of non-communicable diseases (NCDs) such as hypertension, diabetes, cancers, chronic respiratory diseases, and cardiovascular conditions. Health facilities report increasing NCD admissions, frequent stockouts of essential NCD medicines, poor follow-up of chronic patients, and limited integration of NCD data across primary, secondary, and tertiary care. In response, the Ministry of Health has established the National Non-Communicable Disease Intelligence and Analytics Platform (NNCDIAP), integrating hospital EMR data, NCD clinic registers, laboratory information systems, pharmacy/LMIS records, NHIS/insurance claims, community screening data, and UBOS demographic/socio-economic statistics.

You have been hired as a Big Data Analyst to investigate NCD burden, service utilisation, treatment outcomes, and health system readiness using a dataset of at least 10,000 NCD-related records. You may obtain a relevant NCD or chronic-disease dataset from reputable open-data platforms such as Kaggle, WHO Global Health Observatory, World Bank Open Data, or any other credible repository. The dataset must contain sufficient variables to support descriptive analytics, inferential statistics, predictive modelling, machine learning, clustering,

and data visualisation using R.

TASKS:

i. **DESCRIPTIVE ANALYTICS (12 Marks)**

Conduct an exploratory descriptive assessment of key NCD indicators such as patient demographics, disease types (e.g., hypertension, diabetes, cancers), comorbidities, service utilisation (outpatient vs inpatient), medication patterns, and treatment outcomes (e.g., controlled vs uncontrolled, readmissions). Highlight notable disparities across regions, facility levels, gender, or age groups.

ii. **INFERENTIAL ANALYTICS (18 Marks)**

Investigate statistical relationships and differences among selected variables related to NCD risk factors (e.g., age, BMI, smoking, alcohol use), facility characteristics, treatment regimens, and patient outcomes (e.g., control status, complications, readmissions). Apply suitable inferential methods (e.g., correlation, t-tests, ANOVA, chi-square, regression-based tests) to draw evidence-based conclusions.

iii. **PREDICTIVE MODELLING (12 Marks)**

Develop predictive models to estimate key NCD-related outcomes such as probability of hospital readmission, likelihood of poor disease control, risk of NCD complications, or mortality among high-risk patients. Evaluate model performance (e.g., using accuracy, RMSE, AUC, confusion matrix) and explain the significance and interpretation of the most important predictors.

iv. **MACHINE LEARNING & CLUSTERING (8 Marks)**

Apply appropriate machine learning and clustering techniques (e.g., k-means, hierarchical clustering, decision trees, random forests) to identify patient risk segments, disease pattern clusters, or facility performance profiles. Explain how the discovered clusters or models can inform targeted interventions, differentiated care, or resource allocation for NCD management.

v. **DATA VISUALISATION (6 Marks)**

Produce clear, policy-relevant visualisations in R (e.g., bar charts, line graphs, heatmaps, maps, risk-score plots) to communicate NCD trends, geographic or socio-demographic disparities, and system inefficiencies (such as stockouts or high readmission rates). Ensure visualisations are properly labelled, interpretable, and suitable for presentation to Ministry of Health decision-makers.

vi. **POLICY & SYSTEM RECOMMENDATIONS (4 Marks)**

Based on your analytical findings, provide well-justified policy recommendations to strengthen NCD prevention and control in Uganda. Propose specific enhancements to NNCDIAP (e.g., new data elements, integration features, dashboards, alert systems) that would support more effective data-driven planning, monitoring, and clinical decision-making.

Question 7:

QUESTION SEVEN (60 MARKS)

Uganda continues to face escalating environmental challenges including deforestation, wetland encroachment, air and water pollution, waste-management crises, soil degradation, biodiversity loss, and increasing climate-related disasters such as floods, droughts, and landslides. These issues are exacerbated by rapid urbanisation, unregulated industrial activity, population pressure, poor enforcement of environmental policies, and fragmented data systems across agencies such as NEMA, the Ministry of Water and Environment, the Uganda Wildlife Authority (UWA), the National Forestry Authority (NFA), and local governments.

To improve environmental planning and enforcement, the government has developed the National Environmental Intelligence and Monitoring Platform (NEIMP) integrating satellite imagery, air-quality sensor data, hydrological datasets, pollution monitoring logs, land-use/land-cover datasets, waste-management records, forest inventories, biodiversity datasets, weather data, and UBOS demographic information.

You have been contracted as a Big Data Analyst to assess environmental degradation patterns, climate-risk hotspots, pollution levels, conservation performance, and resource-management outcomes using a dataset of at least 10,000 environmental records. You may obtain a relevant environmental dataset from reputable sources such as UNEP, NASA Earthdata, World Bank Open Data, Kaggle, Global Forest Watch, or any credible repository. The dataset must contain adequate variables for descriptive analytics, inferential analysis, predictive modelling, machine learning, clustering, and high-quality R visualisation.

TASKS

i. DESCRIPTIVE ANALYTICS (12 Marks)

Conduct a comprehensive descriptive and exploratory analysis of key environmental indicators such as air-quality levels, water-quality metrics, deforestation patterns, land-use changes, waste-generation rates, biodiversity indices, and climate-risk indicators (e.g., rainfall variability, drought severity). Highlight major disparities across regions, ecosystems, urban vs rural areas, or industrial zones.

ii. INFERENTIAL ANALYTICS (18 Marks)

Apply suitable inferential statistical techniques to examine relationships, group differences, or emerging patterns involving pollution levels, land degradation, biodiversity loss, waste-management practices, climate variables, and population or industrial-activity characteristics. Present statistically supported interpretations that can guide environmental policy, enforcement, and planning.

iii. PREDICTIVE MODELLING (12 Marks)

Develop predictive models for key environmental outcomes such as probability of pollution spikes, likelihood of deforestation, waste-generation levels, water-quality deterioration, or climate-risk severity. Assess model performance using suitable metrics and interpret the

significance of critical predictors within Uganda's environmental-management context.

iv. MACHINE LEARNING & CLUSTERING (8 Marks)

Apply appropriate machine-learning or clustering approaches to identify environmental risk clusters, pollution hotspots, ecological zones with similar degradation patterns, or conservation-priority profiles. Explain how the resulting clusters can support targeted environmental monitoring, enforcement, and sustainable-resource management.

v. DATA VISUALISATION (6 Marks)

Produce high-quality, policy-ready R visualisations (e.g., heatmaps, spatial maps, trend lines, pollution density plots, deforestation-time-series graphs) that effectively communicate environmental risks, disparities, and trends. Briefly justify why each chosen visualisation is appropriate for the insights being communicated.

vi. POLICY & SYSTEM RECOMMENDATIONS (4 Marks)

Drawing from your analytical insights, propose actionable environmental-policy recommendations. Suggest practical improvements to NEIMP for example: real-time alerts, dashboard enhancements, predictive-risk maps, integration with enforcement systems, or automated compliance-monitoring tools to strengthen Uganda's environmental governance and climate resilience.

Question 8:

QUESTION EIGHT (60 MARKS)

Uganda's Directorate of Citizenship and Immigration Control (DCIC) faces increasing operational and service-delivery challenges in the management of passports, visas, work permits, citizenship applications, and border-control processes. Key issues include long passport processing times, increasing application volumes, fraud and identity-related risks, system downtimes, delays in biometric verification, congestion at passport offices, low compliance with online application procedures, and limited integration across systems such as NIRA, Interpol, immigration border systems, and consular services abroad.

To strengthen efficiency and intelligence-led immigration management, the Ministry of Internal Affairs has developed the National Immigration Intelligence and Analytics Platform (NIIAP), integrating passport-application data, NIRA demographic records, e-visa system logs, border-entry/exit records, Interpol watchlists, biometric data, consular service records, and online-payment logs.

You have been hired as a Big Data Analyst to examine passport-service performance, fraud risks, application trends, processing bottlenecks, and compliance patterns using a dataset of at least 10,000 immigration-related records. You may obtain a relevant immigration or public-service dataset from credible repositories such as Kaggle, IOM datasets, Open Government Data portals, or synthetic immigration datasets. The chosen dataset must support descriptive analytics, inferential statistics, predictive modelling, machine learning, clustering, and data visualisation using R.

TASKS

i. DESCRIPTIVE ANALYTICS (12 Marks)

Conduct exploratory descriptive analytics on key immigration indicators such as application volumes, processing times, applicant demographics, passport categories, biometric-capture performance, service-centre workloads, visa/work-permit patterns, and compliance with required steps. Highlight disparities across regions, age groups, passport types, or service centres (e.g., Kampala, Mbarara, Gulu, Mbale, diaspora missions).

ii. INFERENTIAL ANALYTICS (18 Marks)

Using suitable inferential methods, investigate relationships and group differences involving processing times, demographic characteristics, fraud-risk indicators, service-centre efficiency, biometric success rates, or payment/verification patterns. Provide statistically defensible interpretations relevant to immigration service delivery and identity-management operations.

iii. PREDICTIVE MODELLING (12 Marks)

Develop predictive models to estimate key immigration outcomes such as likelihood of delayed passport processing, fraud-risk probability, visa-approval likelihood, customer no-shows, or biometric-verification failure. Evaluate model performance and interpret the importance of key predictors in the context of improving DCIC service efficiency and fraud mitigation.

iv. MACHINE LEARNING & CLUSTERING (8 Marks)

Apply appropriate machine-learning or clustering techniques to identify meaningful applicant segments such as high-risk profiles, common fraud-pattern clusters, workload-based service-centre categories, or travel-behaviour segments (based on border-entry/exit logs). Explain the operational and policy relevance of these clusters for immigration management.

v. DATA VISUALISATION (6 Marks)

Produce clear and decision-ready R visualisations (e.g., processing-time trend plots, regional disparity maps, fraud-risk heatmaps, service-centre workload charts) that effectively communicate patterns, gaps, or emerging risks within immigration data. Justify the choice of each visualisation method.

vi. POLICY & SYSTEM RECOMMENDATIONS (4 Marks)

Based on your analytical findings, provide actionable recommendations to improve passport processing, identity management, digital service compliance, and border-control intelligence. Propose enhancements to NIIAP (e.g., real-time processing dashboards, automated fraud alerts, integration with external systems, predictive-risk scoring tools) to strengthen evidence-driven immigration administration.

Question 9:

QUESTION NINE (60 MARKS)

The rapid growth of social media in Uganda especially platforms such as WhatsApp, TikTok, Facebook, Instagram, X, YouTube, and Snapchat has significantly influenced how learners access information, interact, study, and form academic habits. While social media has

potential benefits such as collaborative learning, quick information sharing, and exposure to global knowledge, it also presents challenges including cognitive overload, distraction, reduced attention span, misinformation, academic procrastination, online peer pressure, and mental-fatigue risks. Yet, education stakeholders lack integrated data systems to understand how social media use affects learners' cognitive processes, academic performance, digital behaviour, and learning outcomes.

To address this gap, the Ministry of Education has developed the Social Media and Cognitive Learning Intelligence Platform (SM-CLIP), integrating LMS logs, student performance data, smartphone usage statistics, social media activity logs (duration, frequency, content types), demographic data, wellbeing/mental-health survey findings, and classroom engagement analytics.

You have been appointed as a Big Data Analyst to examine the relationship between social media use, cognitive processes, and learning outcomes using a dataset of at least 10,000 learner-level records. You may obtain any relevant digital-learning or behavioural dataset from Kaggle, UNESCO, World Bank EdStats, or other reputable sources, provided it contains variables that support the required analytics.

TASKS

i. DESCRIPTIVE ANALYTICS (12 Marks)

Perform descriptive and exploratory analytics on key indicators such as social-media usage frequency, time spent on various platforms, types of content consumed (educational vs entertainment), multitasking patterns, attention span indicators, cognitive-engagement metrics, and academic performance. Identify notable disparities across age groups, school types, gender, or regions.

ii. INFERENTIAL ANALYTICS (18 Marks)

Using appropriate inferential techniques, examine relationships and differences among social-media variables (e.g., usage duration, platform type, content exposure) and cognitive-process indicators such as attention, memory performance, task-switching ability, comprehension depth, and overall academic outcomes. Draw statistically supported conclusions relevant to behaviour, cognition, and learning.

iii. PREDICTIVE MODELLING (12 Marks)

Develop predictive models to estimate outcomes such as likelihood of academic decline, cognitive overload, reduced attention span, improved engagement through educational content, or probability of procrastination. Evaluate model performance and interpret predictors in a policy-relevant, learner-focused context.

iv. MACHINE LEARNING & CLUSTERING (8 Marks)

Apply clustering or machine-learning methods to identify meaningful learner segments, such as heavy entertainment-content users, educational-content consumers, multitaskers, high-risk distraction groups, or cognitively resilient digital learners. Explain the pedagogical and policy value of the identified clusters.

v. DATA VISUALISATION (6 Marks)

Produce clear visualisations in R such as behavioural-pattern graphs, cognitive-risk heatmaps, attention-span trend charts, or platform-usage profiles—aimed at communicating critical insights to decision-makers, psychologists, teachers, and curriculum planners.

vi. POLICY & SYSTEM RECOMMENDATIONS (4 Marks)

Based on your findings, propose evidence-based recommendations to improve digital-wellbeing policies, enhance safe and productive social-media use, support cognitive-health interventions, and guide the integration of social-media analytics into SM-CLIP for proactive monitoring of learners' cognitive processes and academic risk factors.

Question 10:

QUESTION TEN (60 MARKS)

Uganda continues to experience rising air pollution levels, especially in major urban centres such as Kampala, Jinja, Mbarara, and Gulu. Key drivers include vehicle emissions, industrial activity, open burning of waste, biomass fuel use, unregulated construction, dust from unpaved roads, and rapid urbanisation. Air quality challenges increasingly contribute to respiratory illnesses, cardiovascular complications, reduced productivity, and worsening environmental health outcomes. However, monitoring systems remain fragmented across agencies, with limited integration of meteorological data, sensor readings, satellite observations, traffic-flow logs, and health data.

To strengthen evidence-based environmental management, the Ministry of Water and Environment has developed the National Air Quality Intelligence Platform (NAQIP), integrating low-cost sensor data, satellite-derived aerosol measurements, meteorological variables, industrial emissions logs, traffic density data, health facility respiratory-case records, and UBOS demographic data.

You have been appointed as a Big Data Analyst to assess air-pollution patterns, identify hotspots, understand exposure risks, and generate actionable insights using a dataset of at least 10,000 air-quality and environmental-health records. You may obtain relevant datasets from sources such as NASA Earthdata, OpenAQ, World Bank Climate Data, UNEP datasets, Kaggle, or any other reputable repository.

TASKS

i. DESCRIPTIVE ANALYTICS (12 Marks)

Conduct an exploratory descriptive analysis on key air-quality indicators such as PM2.5, PM10, NO_2, CO, O_3, SO_2 levels, temperature, humidity, wind patterns, traffic density, and health-facility respiratory case counts. Identify pollution disparities across regions, urban vs rural settings, time of day, seasons, or industrial zones.

ii. INFERENTIAL ANALYTICS (18 Marks)

Using appropriate inferential techniques, examine relationships and group differences involving pollution levels, meteorological conditions, population density, traffic patterns, industrial activity, or respiratory health outcomes. Provide statistically supported

interpretations to guide air-quality management and environmental-health policy.

iii. PREDICTIVE MODELLING (12 Marks)

Develop predictive models to estimate outcomes such as future PM2.5 levels, likelihood of pollution spikes, respiratory-disease risk, or high-exposure zones. Evaluate the predictive accuracy of the models and interpret key predictors relevant to environmental planning and public-health protection.

iv. MACHINE LEARNING & CLUSTERING (8 Marks)

Apply relevant machine-learning or clustering methods to identify pollution hotspots, temporal pollution pattern groups, meteorological-pollution behaviour clusters, or health-risk exposure profiles. Explain the operational and policy relevance of the clustering results.

v. DATA VISUALISATION (6 Marks)

Produce clear, decision-ready visualisations in R such as pollution-trend graphs, geospatial pollution-hotspot maps, pollution-health correlation plots, or diurnal/seasonal variation charts. Justify your choice of visualisations in relation to the insights they convey.

vi. POLICY & SYSTEM RECOMMENDATIONS (4 Marks)

Based on your analytical findings, propose actionable recommendations to improve air-quality monitoring, pollution control, and public-health protection. Suggest enhancements for NAQIP—such as real-time alerts, automated hotspot detection, integration with city-traffic systems, or predictive early-warning dashboards—to support evidence-driven environmental governance.

Question 11:

QUESTION ELEVEN (60 MARKS)

Uganda continues to experience persistent challenges in election administration, transparency, and data-driven decision-making during national and local elections. Key concerns include inconsistent voter registers, delayed results transmission, inadequate polling-station logistics, uneven distribution of electoral materials, limited real-time monitoring, election-related violence reports, misinformation on social media, and logistical inefficiencies in remote districts. Although institutions such as the Electoral Commission (EC), National Identification and Registration Authority (NIRA), Uganda Communications Commission (UCC), security agencies, and civil-society organisations generate large volumes of election-related data, these datasets remain fragmented and poorly integrated.

To strengthen evidence-based electoral management, the Electoral Commission has developed the National Electoral Intelligence Platform (NEIP). NEIP integrates polling-station data, biometric verification logs, polling-station turnout records, NIRA demographic datasets, results-transmission timestamps, electoral-materials distribution logs, security-incident reports, social-media misinformation indicators, and historical election results.

You have been appointed as a Big Data Analyst to assess electoral patterns, detect

anomalies, evaluate turnout disparities, and generate actionable insights using a dataset of at least 10,000 election-related records.

I. DESCRIPTIVE ANALYTICS (12 MARKS)

Conduct an exploratory descriptive analysis on key electoral indicators such as voter turnout, biometric verification success rates, polling-station opening/closing times, results-transmission delays, distribution of polling materials, demographic participation patterns, and election-day incident reports. Identify turnout disparities across districts, age groups, gender, rural vs urban settings, polling-station categories, or historical voting patterns.

II. INFERENTIAL ANALYTICS (18 MARKS)

Using relevant inferential statistical techniques, examine relationships and group differences among variables such as turnout, demographic characteristics, verification success, polling-station congestion, results-transmission delays, or incident occurrence. Provide statistically supported interpretations to guide electoral-process strengthening, logistical planning, and voter-participation improvement.

III. PREDICTIVE MODELLING (12 MARKS)

Develop predictive models to estimate outcomes such as polling-station turnout, likelihood of verification failure, probability of results-transmission delays, or risk of election-day incidents. Evaluate model accuracy and interpret predictors relevant to election-administration planning, voter-education strategies, and early-warning decision support.

IV. MACHINE LEARNING & CLUSTERING (8 MARKS)

Apply suitable machine-learning or clustering methods to identify groups such as high-risk polling stations, turnout-behaviour clusters, demographic participation profiles, or patterns of misinformation spread. Explain the operational relevance of your clustering results for real-time monitoring, resource allocation, or electoral-integrity interventions.

V. DATA VISUALISATION (6 MARKS)

Produce clear and decision-focused visualisations in R, such as turnout-trend graphs, geospatial maps of results-transmission delays, demographic participation heatmaps, polling-station risk clusters, or time-series plots of verification performance. Justify your visualisation choices based on clarity, insight, and relevance to election management.

VI. POLICY & SYSTEM RECOMMENDATIONS (4 MARKS)

Based on your analytical findings, propose practical recommendations to improve election management, voter participation, election-technology reliability, and results-transmission transparency. Suggest enhancements for NEIP—such as real-time turnout monitoring

dashboards, automated anomaly detection, integration with digital IDs, or predictive early-warning alerts—to support evidence-driven electoral governance in Uganda.

Question 12:

QUESTION TWELVE(60 MARKS)

Uganda continues to face challenges in higher education planning, particularly in the alignment between university programme intakes and labour-market demands. Many universities admit large numbers of students into programmes with limited market absorption, while critical sectors such as agricultural technology, data science, engineering, and health technologies remain understaffed. The absence of coordinated national skills planning, outdated curriculum structures, limited employer engagement, weak career-guidance systems, and fragmented labour-market data contribute to increasing unemployment among graduates.

To address this growing concern, the Ministry of Education and Sports (MoES), together with the National Council for Higher Education (NCHE), has developed the National Skills & Labour-Market Intelligence Platform (NSLMIP). The platform integrates datasets from university admission records, graduation statistics, employer skills-demand surveys, UBOS labour-force data, internship placement logs, national curriculum frameworks, and job-market analytics from recruitment portals.

You have been appointed as a Big Data Analyst to analyse university intake patterns, identify skills mismatches, assess employability risks, and generate actionable insights using a dataset of at least 10,000 higher-education and labour-market records.

I. DESCRIPTIVE ANALYTICS (12 MARKS)

Conduct an exploratory analysis on key indicators such as programme intake volumes, graduation rates, internship placement rates, employer-requested skills, unemployment levels by programme, sector-specific labour demand, and regional programme distribution. Identify disparities across universities, academic disciplines, gender, regions, and public vs private institutions.

II. INFERENTIAL ANALYTICS (18 MARKS)

Using appropriate inferential statistical methods, examine relationships and group differences between variables such as programme intake size, curriculum relevance, labour-market demand levels, graduate employment outcomes, employer skill requirements, and unemployment duration. Provide statistically supported interpretations to guide policy and academic reforms.

III. PREDICTIVE MODELLING (12 MARKS)

Develop predictive models to estimate outcomes such as likelihood of graduate unemployment, probability of programme oversubscription, future labour-market shortages in critical sectors, or projected demand for specific skills. Evaluate model accuracy and interpret key predictors relevant for national skills planning and university admissions policy.

IV. MACHINE LEARNING & CLUSTERING (8 MARKS)

Apply clustering or other machine-learning methods to identify skills-mismatch clusters, high-risk academic programmes, university intake behaviour patterns, or student-employability profiles. Explain the practical relevance of the clusters for curriculum redesign, admissions planning, and national workforce development.

V. DATA VISUALISATION (6 MARKS)

Produce decision-ready visualisations in R such as programme-intake trend graphs, employability heatmaps, sectoral labour-demand maps, skills-mismatch dashboards, or unemployment-duration distributions. Justify your visualisation choices in relation to insights required for policy and university management.

VI. POLICY & SYSTEM RECOMMENDATIONS (4 MARKS)

Based on your analytical findings, propose actionable recommendations to improve university admissions planning, enhance curriculum alignment with labour-market needs, and reduce graduate unemployment. Suggest enhancements for NSLMIP—such as automated skills-gap alerts, admissions-quota optimisation tools, employer–university collaboration dashboards, or predictive workforce-planning modules—to support evidence-driven decision-making.