

**Institute of Primate Research**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **SOP No.** | **Issue Number** | **Issue Date** | **Revision Status** | **Revision Date** |
| **SOP/KIPRE/RPD/DSAS/3.1.76** | **Version 01** | **October 2025** | **-** | **-** |

**STANDARD OPERATING PROCEDURE (SOP) DOCUMENT**

**Predictive Modelling and Ensemble Modelling**

|  |  |  |  |
| --- | --- | --- | --- |
| **Approvals** |  |  |  |
|  | **Name** | **Signature** | **Date** |
| **Developed by:** | \_Patrick Waweru Mwaura\_ | **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** | **\_6th October; 2025\_** |
|  | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** | **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** |
|  | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** | **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** |
| **Reviewed by:** | **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** | **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** | **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** |
| **Approved by:** | **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** | **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** | **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |

**Table of Contents**

[1. PURPOSE 4](#_Toc144316958)

[2. SCOPE 4](#_Toc144316959)

[3. PERSONS RESPONSIBLE: 4](#_Toc144316960)

[4. FREQUENCY 4](#_Toc144316961)

[5. MATERIALS 4](#_Toc144316962)

[6. PROCEDURE 4](#_Toc144316963)

[7. REFERENCES 7](#_Toc144316964)

# 

# PURPOSE

To establish a standardized framework for developing, validating, and applying predictive and ensemble models in DS&AS-supported research, ensuring accuracy, transparency, and compliance with ethical and regulatory standards.

# SCOPE

Covers all DS&AS projects involving statistical or machine learning predictive modelling, including epidemiology, biomedical outcomes, ecological monitoring, and conservation forecasting.

# PERSONS RESPONSIBLE:

* **Data Scientist / Biostatistician:** Designs and implements models.
* **Computational Biologist (if genomic/proteomic data):** Applies specialized methods.
* **Head of DS&AS:** Reviews and approves modelling frameworks.

# FREQUENCY

* Mandatory **before deployment** of any predictive model.
* Model re-validation **when new data or methods** become available.

# MATERIALS

* Statistical software (R, SAS, Python/Scikit-learn, TensorFlow, PyTorch).
* Model validation datasets.
* Documentation templates (model specification, assumptions, metrics).

# PROCEDURE

1. **Model Selection:** Identify candidate models (regression, decision trees, random forests, boosting, and deep learning).
2. **Data Preparation:** Clean, partition (train/validation/test), and preprocess datasets.
3. **Model Development:** Train models using cross-validation; tune hyperparameters.
4. **Validation:** Assess performance (AUC, RMSE, calibration plots, sensitivity analyses).
5. **Ensemble Modelling:** Combine models using bagging, boosting, or stacking where appropriate.
6. **Documentation:** Archive all code, parameters, and validation results.
7. **Deployment:** Deploy validated model with monitoring mechanisms.

# REFERENCES

## **SOP 15: Predictive Modelling and Ensemble Modelling**

### 1. PURPOSE

### 2. SCOPE

### 3. PERSONS RESPONSIBLE

### 4. FREQUENCY

### 5. MATERIALS

### 6. PROCEDURE

## **SOP 16: Handling Large Datasets and Trend Detection**

### 1. PURPOSE

To standardise DS&AS methods for managing large datasets and detecting temporal, spatial, and epidemiological trends.

### 2. SCOPE

Applies to DS&AS projects involving large-scale biomedical, ecological, genomic, or public health data.

### 3. PERSONS RESPONSIBLE

* **Data Engineer:** Manages big data storage and access pipelines.
* **Data Scientist:** Conducts statistical and trend analysis.
* **Head of DS&AS:** Ensures compliance and efficiency.

### 4. FREQUENCY

* Applied continuously for projects requiring high-volume or real-time analytics.
* Reviewed **annually** for optimisation and scalability.

### 5. MATERIALS

* Big data tools (Hadoop, Spark, SQL, NoSQL, PostgreSQL).
* Trend detection tools (time-series models, anomaly detection, geospatial analysis).
* Data visualisation dashboards (R Shiny, Tableau, PowerBI).

### 6. PROCEDURE

1. **Data Ingestion:** Import datasets into scalable storage (SQL/NoSQL databases, distributed systems).
2. **Preprocessing:** Apply automated cleaning, deduplication, and standardisation.
3. **Trend Analysis:** Apply statistical methods (ARIMA, Cox models, GAMs) and ML-based detection (anomaly detection, clustering).
4. **Visualisation:** Generate dashboards for real-time monitoring of patterns and anomalies.
5. **Archiving:** Store processed datasets and scripts in the central repository.
6. **Review:** Conduct annual scalability and performance audits.

## **SOP 17: Reproducible Coding Practices (Git, R Markdown, Jupyter, etc.)**

### 1. PURPOSE

To ensure all DS&AS coding practices are transparent, reproducible, version-controlled, and auditable.

### 2. SCOPE

Applies to all coding activities in DS&AS projects, including statistical analysis, machine learning, bioinformatics pipelines, and reporting.

### 3. PERSONS RESPONSIBLE

* **All DS&AS Analysts and Data Scientists:** Follow reproducible coding practices.
* **Data Engineer:** Maintains version-control repositories.
* **Head of DS&AS:** Monitors compliance with reproducibility standards.

### 4. FREQUENCY

* Applied **throughout project lifecycle**.
* **Quarterly audits** of repositories for compliance.

### 5. MATERIALS

* Git/GitHub/GitLab for version control.
* R Markdown, Jupyter Notebooks, Quarto for documentation.
* Institutional coding standards and templates.

### 6. PROCEDURE

1. **Repository Setup:** Create project repository in Git (institutional or GitHub Enterprise).
2. **Version Control:** Commit all scripts with meaningful messages; use branching for feature development.
3. **Reproducible Documentation:** Use R Markdown/Jupyter/Quarto for analyses, embedding code and results together.
4. **Collaboration:** Use pull requests and peer code review before merging.
5. **Archiving:** Tag final project versions; store outputs and code in DS&AS repository.
6. **Audit:** Quarterly review of repositories for compliance with reproducibility standards.