

**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 5](#_Toc144316962)

[6. PROCEDURE 5](#_Toc144316963)

[7. REFERENCES 6](#_Toc144316964)

# 

# PURPOSE

To provide a standardized framework for the development, validation, deployment, and documentation of predictive and ensemble models within DS&AS-supported research projects, ensuring:

* Methodological rigor and reproducibility in alignment with institutional statistical and computational standards (SOPs 1, 4, 14).
* Compliance with ethical, regulatory, and data governance requirements (SOPs 2, 6–9).
* Transparent reporting, version control, and auditable records of all modelling activities.
* Continuous evaluation and updating of models based on new data, methods, or performance assessments.

# SCOPE

Applies to all DS&AS-supported research projects that involve the development, validation, and application of predictive and ensemble models, including but not limited to:

* Epidemiological modelling for disease surveillance and control.
* Predictive analytics for biomedical outcomes and translational research.
* Ecological and conservation forecasting.
* Genomic and proteomic data-driven predictions.

This SOP governs all stages of modelling, from data preparation and model selection to validation, deployment, and documentation, in accordance with institutional policies (SOP 1), ethical and regulatory standards (SOP 2), statistical analysis plans (SOP 4), and computational tool validation procedures (SOP 14).

# PERSONS RESPONSIBLE:

* **Data Scientist / Biostatistician:** Leads model development, ensures adherence to statistical principles (SOP 3, SOP 4), and implements predictive/ensemble models.
* **Computational Biologist / Bioinformatician (if genomic/proteomic data involved):** Applies domain-specific modelling methods, ensures compliance with genome/proteome data management (SOP 12, SOP 13).
* **Head of DS&AS:** Reviews and approves model specifications, validates documentation and compliance with institutional policies (SOP 1), ethical standards (SOP 2), and computational tool validation (SOP 14).

# FREQUENCY

* **Pre-deployment:** All predictive and ensemble models must undergo validation before being used for analysis or decision-making.
* **Re-validation:** Models must be reviewed and re-validated whenever:
* New datasets are added or existing data significantly updated.
* Changes to modelling methods, algorithms, or software versions occur.
* Regulatory or ethical requirements impacting model use are updated (see SOP 2, SOP 14).
* **Periodic Review:** At least annually, DS&AS conducts a review of active models to ensure continued accuracy and compliance.

# MATERIALS

* **Statistical and Machine Learning Software:** R (caret, mlr3, tidymodels), SAS, Python (scikit-learn, TensorFlow, PyTorch, XGBoost, LightGBM).
* **Validation Datasets:** Independent datasets for model training, validation, and testing; benchmark datasets where applicable.
* **Documentation Templates:** Standardized templates for model specifications, assumptions, performance metrics, and versioning logs.
* **Computational Resources:** HPC/Cloud infrastructure for training and deployment of models.
* **Version Control Systems:** Git/GitHub/GitLab repositories for reproducible code management.
* **Reporting Tools:** Visualization tools (ggplot2, matplotlib, seaborn) and dashboards for communicating model performance.

# PROCEDURE

1. **Model Selection**:

* Identify candidate models appropriate for the data and outcome type (e.g., linear/logistic regression, decision trees, random forests, gradient boosting, neural networks).
* Consider interpretability, computational resources, and regulatory requirements.

1. **Data Preparation**:

* Clean datasets and handle missing values.
* Partition data into training, validation, and test sets.
* Apply preprocessing (normalization, feature encoding, feature engineering) consistently across partitions.

1. **Model Development**:

* Train models on the training set using cross-validation to prevent overfitting.
* Tune hyperparameters systematically (grid search, random search, Bayesian optimization).
* Document assumptions, parameter choices, and rationale for model selection.

1. **Validation**:

* Evaluate model performance on validation and test sets using metrics appropriate to the problem (AUC, RMSE, accuracy, sensitivity/specificity, calibration plots).
* Conduct sensitivity analyses and assess generalizability.
* Record performance results for audit and reproducibility.

1. **Ensemble Modelling**:

* Combine multiple models where appropriate using bagging, boosting, or stacking.
* Validate ensemble performance against individual models to ensure improvement.

1. **Documentation**:

* Archive all code, scripts, parameter settings, datasets used for training/validation, and results.
* Maintain version-controlled repositories (Git/GitHub/GitLab) for reproducibility.

1. **Deployment and Monitoring**:

* Deploy validated model in production or for research use with controlled access.
* Establish monitoring mechanisms to track model performance over time and trigger re-validation when new data or drift is detected.

# REFERENCES

1. Kuhn, M., & Johnson, K. (2013). Applied Predictive Modeling. Springer.
2. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.
3. Wilkinson, M.D. et al., (2016). FAIR Guiding Principles for scientific data management and stewardship. Scientific Data.
4. Kenya Data Protection Act (2019) – ensuring ethical use of personal or sensitive data (linked to SOP 2: Alignment with Institutional and National Regulations; SOP 6: Data Access and Authentication; SOP 9: Data Sharing and Anonymisation).
5. KIPRE Institutional Data Governance and Software Development Guidelines (linked to SOP 1: Policies & Strategies; SOP 7: Data Storage, Backup, Encryption, and Disaster Recovery; SOP 8: Database and Workflow Management).
6. Sandve, G.K., et al., (2013). Ten Simple Rules for Reproducible Computational Research. PLoS Comput Biol (linked to SOP 14: Development and Validation of Computational Tools; SOP 13: Bioinformatics Pipelines).
7. Git/GitHub/GitLab documentation – for version-controlled code and reproducibility (linked to SOP 4: Statistical Analysis Plans; SOP 13: Bioinformatics Pipelines; SOP 14: Development and Validation of Computational Tools).

# APPENDIX: Forms and Templates for Predictive and Ensemble Modelling

**A1. Predictive Model Specification Form**

| **Field** | **Description** |
| --- | --- |
| Project Title | Name of the research project |
| PI / Research Team | Names and roles |
| Data Source | Dataset(s) used for modelling |
| Outcome Variable(s) | Dependent variables to predict |
| Predictor Variables | Independent variables/features |
| Model Type | Regression, decision tree, random forest, boosting, deep learning, etc. |
| Assumptions | List model assumptions and data considerations |
| Preprocessing Steps | Scaling, normalization, missing data handling |
| Training/Validation Split | Method and proportion (e.g., 70/30, cross-validation) |
| Hyperparameters | Initial values or tuning ranges |
| Expected Metrics | Performance metrics to evaluate (AUC, RMSE, sensitivity, specificity, etc.) |
| Date & Version | Version of the model specification |

**A2. Model Validation Report Template**

| **Field** | **Description** |
| --- | --- |
| Model Name & Version | Name and version number of the model |
| Validation Dataset | Dataset used for validation |
| Metrics | Quantitative results (AUC, RMSE, MAE, calibration, etc.) |
| Cross-Validation Results | Summary of folds, mean, SD |
| Sensitivity Analyses | Analyses conducted to check robustness |
| Limitations | Known limitations of the model |
| Reviewer | Name of internal reviewer |
| Date | Date of validation completion |

**A3. Ensemble Modelling Record**

| **Field** | **Description** |
| --- | --- |
| Ensemble Name & Version | Name of ensemble model |
| Constituent Models | List of individual models included in ensemble |
| Combination Method | Bagging, boosting, stacking, weighted averaging |
| Validation Metrics | Performance of ensemble vs. individual models |
| Notes | Observations on model performance or adjustments made |
| Date & Version | Versioning information |

**A4. Deployment & Monitoring Checklist**

| **Field** | **Description** |
| --- | --- |
| Model / Ensemble Version | Name and version number |
| Deployment Platform | R, Python, Shiny, web API, HPC/Cloud |
| Monitoring Plan | Frequency and type of monitoring (accuracy drift, input data changes, logs) |
| Responsible Staff | Name of person/team monitoring the model |
| Documentation Location | Repository or folder for scripts, logs, results |
| Approval | Head of DS&AS sign-off for deployment |