

**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 4](#_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, and 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