Data-Driven Vaccination:

Enhancing Predictive Models for Public Health

<u>Team 4:</u>

Svetlana Lukina,
Oleg Rakhmatullin,
Dmitrii Kornienko.

10/24/2024 Skoltech

INTRODUCTION

Vaccination is a key element of public health aimed at preventing the spread of infectious diseases.

Despite the availability of vaccines, there are a significant number of people who refuse vaccination due to various factors such as distrust of medical institutions, the influence of public opinion and personal beliefs [1-3].

OUR CORE AIM

- Analyze vaccination data and identify significant patterns that affect the intention to get vaccinated.
- Develop and optimize machine learning models to predict vaccination intentions.
- To propose recommendations for improving vaccination programs based on the results obtained.

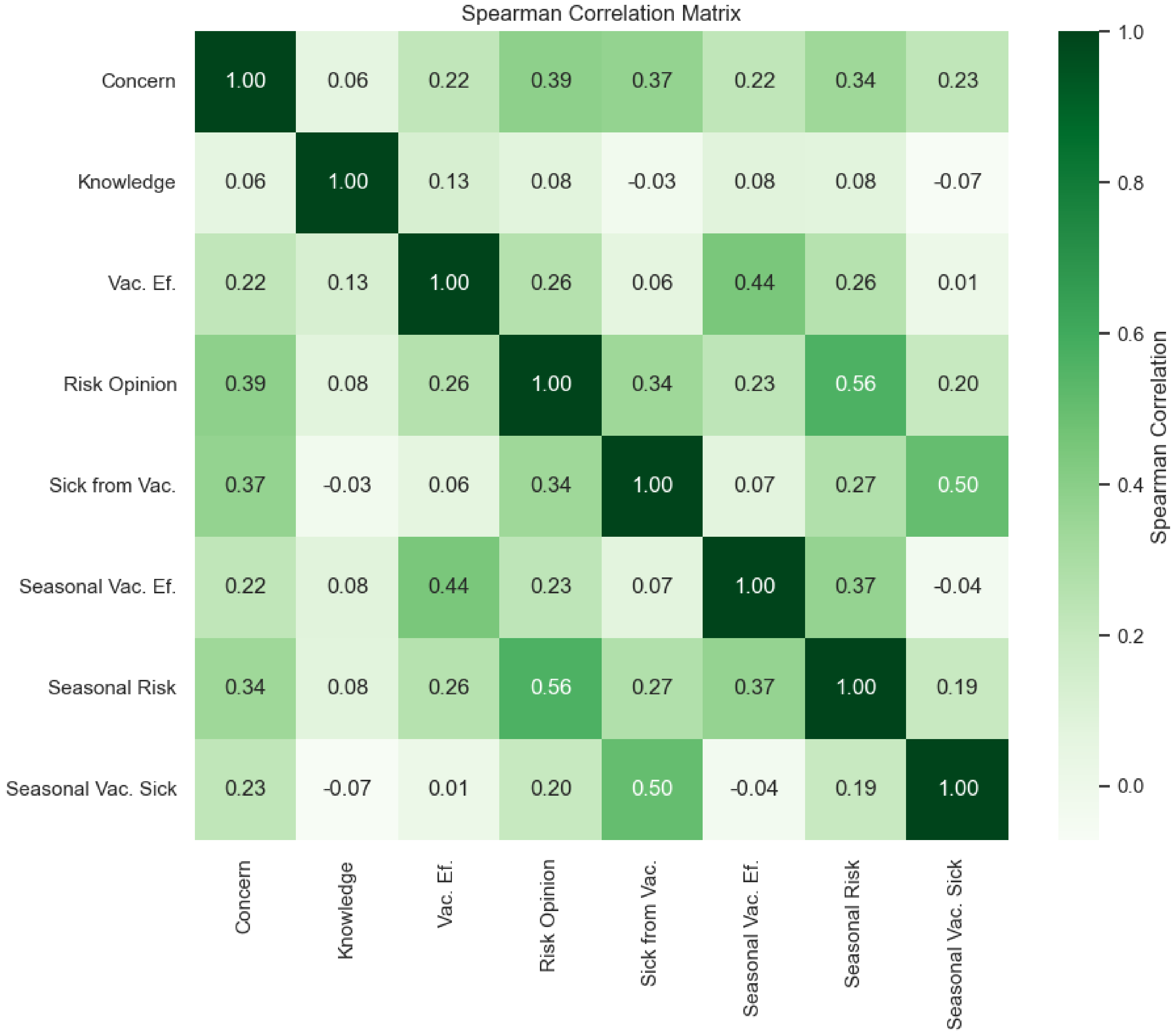
1.Seasonal Influenza Vaccine Impact on Pandemic H1N1 Vaccine Efficacy / Lee, R. U., Phillips, C. J., & Faix, D. J
2. Defining the root cause of reduced H1N1 live attenuated influenza vaccine effectiveness: low viral fitness leads to inter-strain competition / Dibben, O., Crowe, J., Cooper, S., et al.
3.Influenza Vaccine Effectiveness: New Insights and Challenges / Ainslie, K. E. C., Shi, M., Haber, M., & Orenstein, W. A.

DATA ANALYSIS

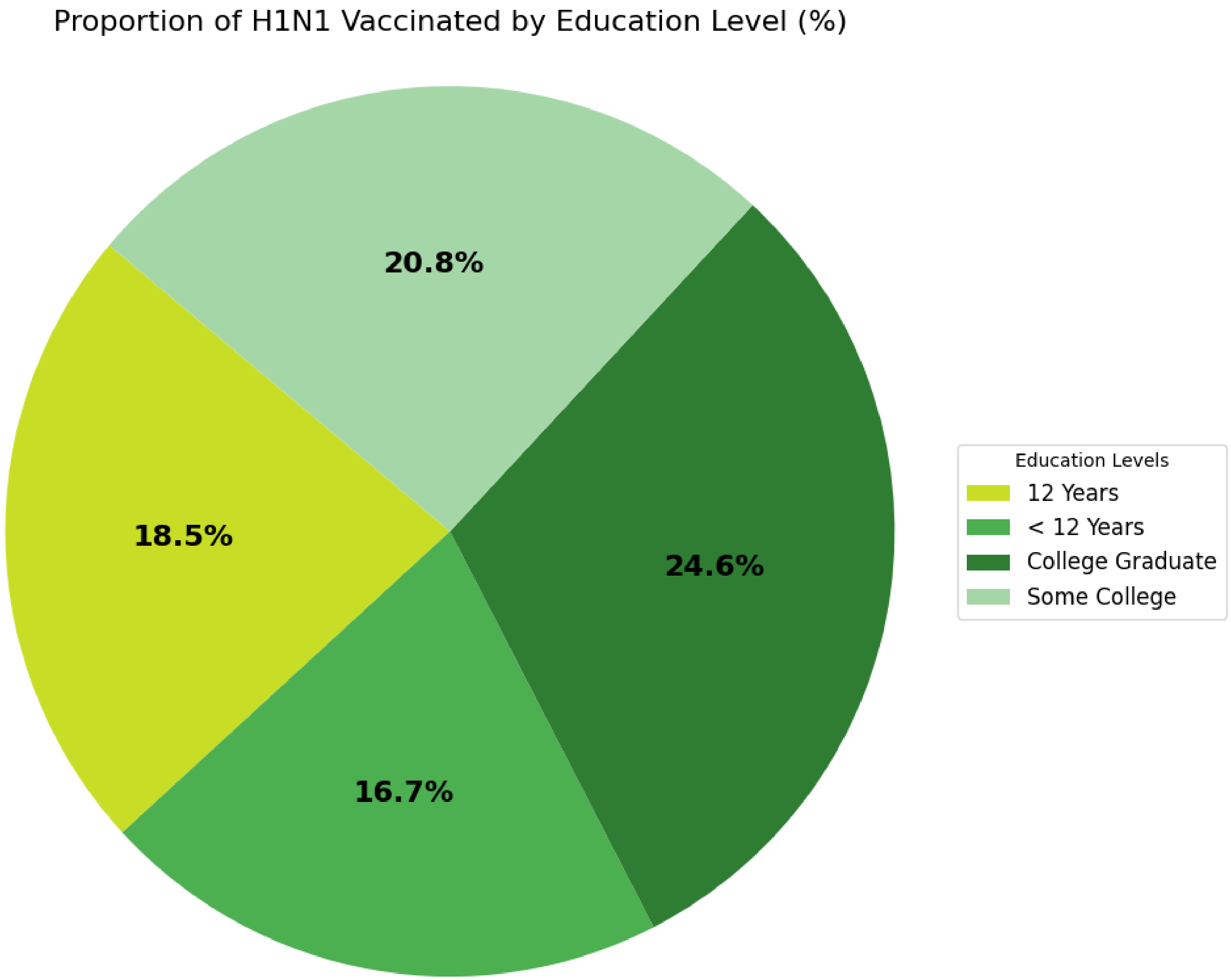
Two target variables:

h1n1_vaccine - Whether respondent received H1N1 flu vaccine.

seasonal_vaccine - Whether respondent received seasonal flu vaccine.

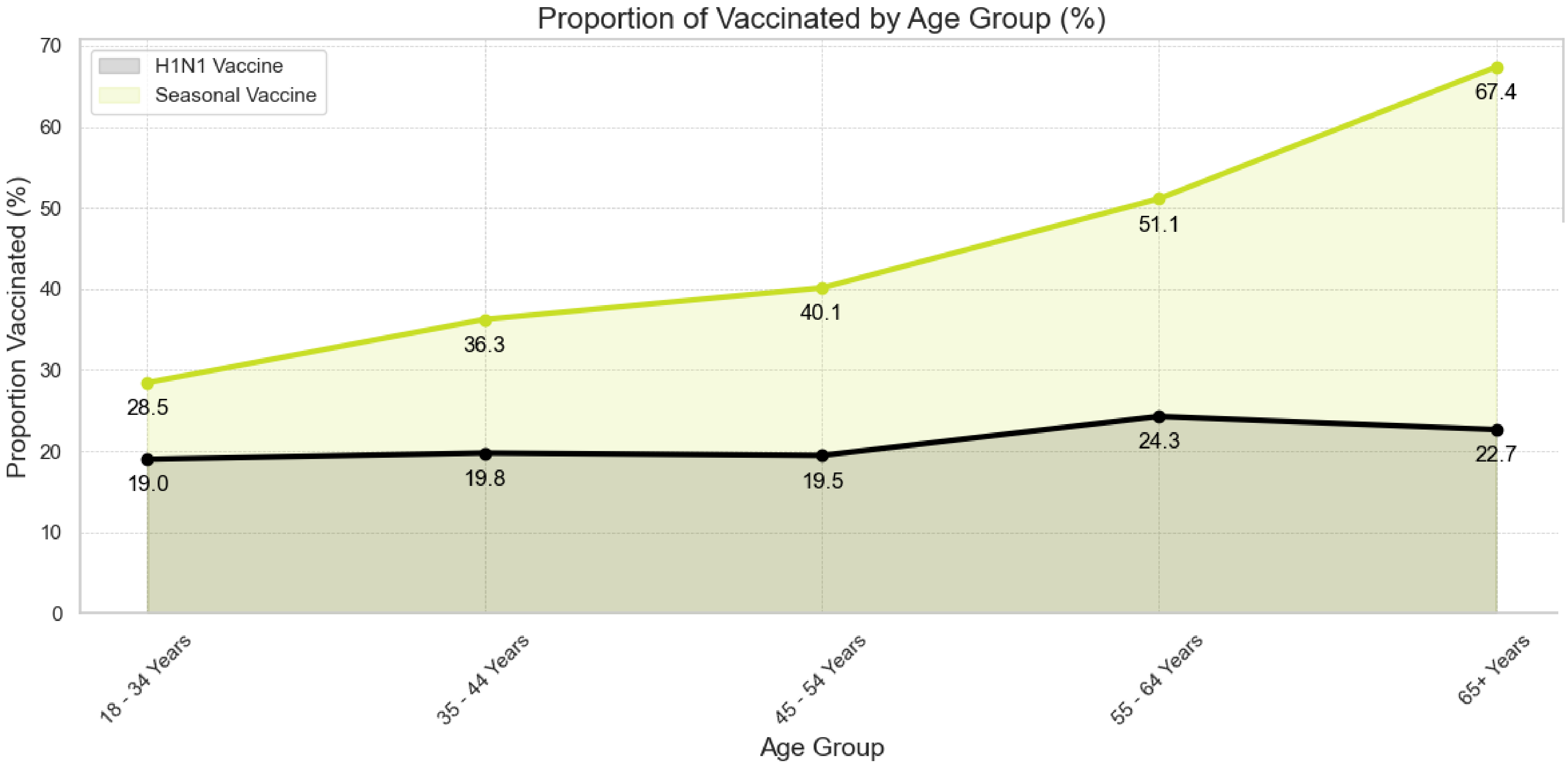


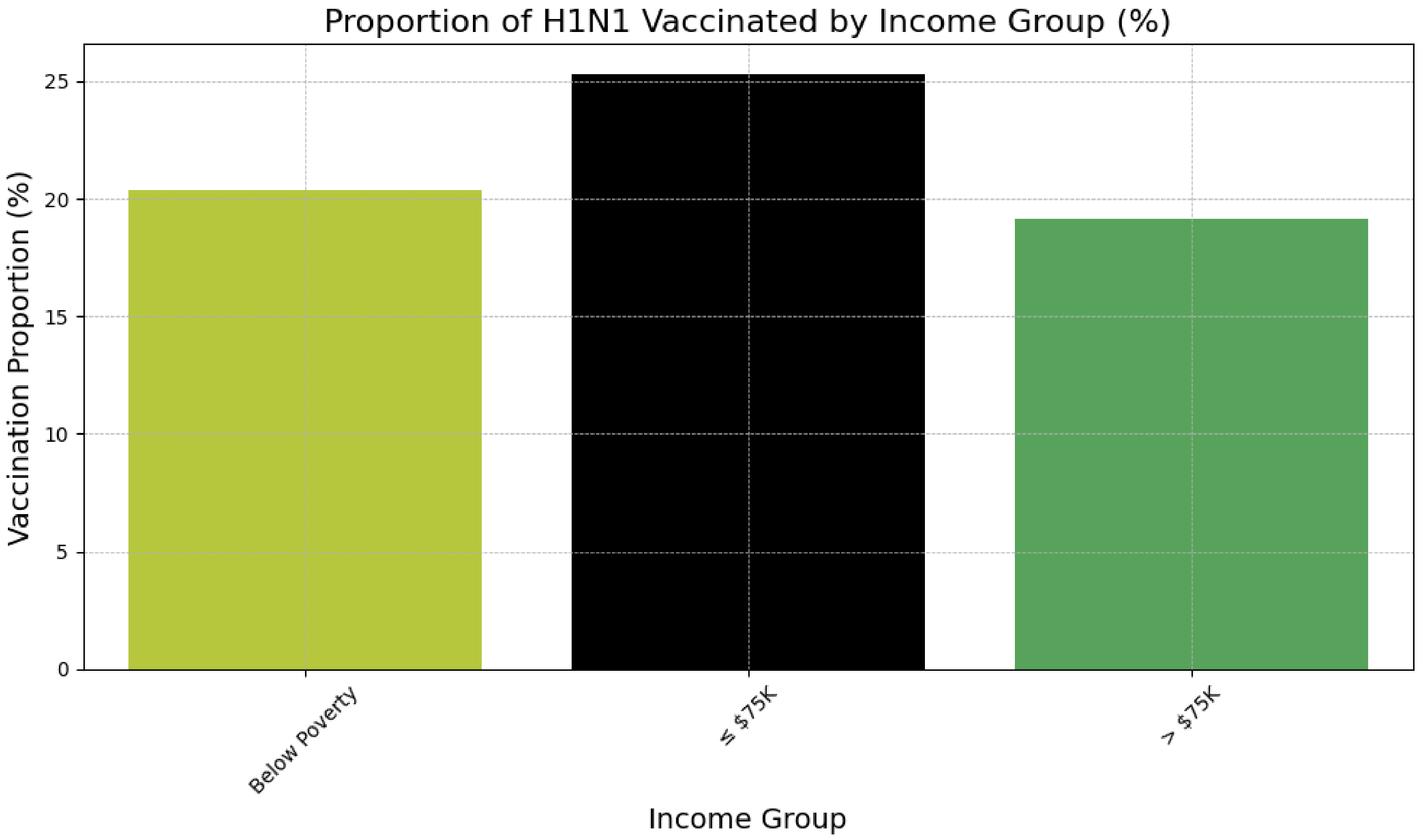
	Concern	Knowledge	Vac. Ef.	Risk Opinion	Sick from Vac.	Seasonal Vac. Ef.	Seasonal Risk	Seasonal Vac. Sick
Concern	1.00	0.06	0.22	0.39	0.37	0.22	0.34	0.23
Knowledge	0.06	1.00	0.13	0.08	-0.03	0.08	0.08	-0.07
Vac. Ef.	0.22	0.13	1.00	0.26	0.06	0.44	0.26	0.01
Risk Opinion	0.39	0.08	0.26	1.00	0.34	0.23	0.56	0.20
Sick from Vac.	0.37	-0.03	0.06	0.34	1.00	0.07	0.27	0.50
Seasonal Vac. Ef.	0.22	0.08	0.44	0.23	0.07	1.00	0.37	-0.04
Seasonal Risk	0.34	0.08	0.26	0.56	0.27	0.37	1.00	0.19
Seasonal Vac. Sick	0.23	-0.07	0.01	0.20	0.50	-0.04	0.19	1.00



Skoltech

DATA ANALYSIS





MODEL SELECTION

Logistic Regression

Random forest

XGBoost

Why did we choose it?

Logistic Regression

1. **Simplicity and Interpretability**: Logistic regression allows for easy interpretation of results, making it suitable for initial analysis.
2. **Effectiveness for Binary Classification:** It is well-suited for tasks where the target variable has two classes, making it a solid baseline for binary classification tasks.
3. **Fast Training:** The model trains quickly on small to medium-sized datasets.

Random forest

1. **Avoidance of local minima:** it is less likely to get stuck in local minima, providing a more robust and stable solution compared to individual decision trees.
2. **Handles missing data:** The algorithm can work effectively with datasets containing missing values without imputation.
3. **High accuracy:** Random Forest generally provides more accurate predictions by aggregating multiple decision trees.

XGBoost

1. **High Performance**: XGBoost often outperforms other algorithms due to its use of gradient boosting.
2. **Handling Large Datasets**: It works efficiently with large datasets and can automatically model interactions between features.
3. **Regularization**: It includes regularization techniques, which robust to overfitting.

Expected Outcomes

Logistic Regression

I. **High Accuracy in Class Probability Predictions**: Especially effective if the data is linearly separable.

II. **Easy Interpretation:** It provides straightforward interpretation of the results and allows easy identification of important features

Random forest

I. **More stable predictions:** By training multiple independent trees, it more stable and reliable predictions, even in the presence of noisy or complex data.

II. **Improved performance on non-linear data:** Random Forest can handle complex, non-linear relationships between features, making it a robust model across various datasets and tasks.

XGBoost

I. **Expected High Accuracy:** XGBoost can deliver high accuracy due to its advanced optimization techniques and resistance to overfitting, thanks to built-in regularization methods..

II. **Ability to Handle Large Volumes:** It excels in managing large datasets and modeling complex, non-linear interactions between features

HYPERPARAMETER TUNING

For tinning we are using **Grid Search.** It is a method that that tests all possible combinations hyperparameters to find the optimal configuration for the model (using performance parameters, e.g. roc_auc, f1)

```
logreg_param_grid = [
    {
        'C': np.logspace(-5, 5, endpoint=True, num=31),
        'penalty': ['l1', 'l2'],
        'solver': ['liblinear'] #optimization algorithm
    },
    {
        'C': np.logspace(-5, 5, endpoint=True, num=31),
        'penalty': ['l2'],
        'solver': ['lbfgs']
    }
]

logreg_grid_search = GridSearchCV(
    estimator=logreg_model,
    param_grid=logreg_param_grid,
    scoring='roc_auc',
    cv=skf,
    verbose=1
)
```

Best Parameters for Logistic Regression:

C = (-5,5), penalty: ['l1', 'l2’]; 'solver': ['liblinear’], max_iter=5000, cv=skf, scoring='roc_auc'

```
rf_model = RandomForestClassifier(random_state=314)

rf_param_grid = {
    'n_estimators': [50, 75, 100, 125, 150, 175, 200, 225],
    'max_depth': [3, 5, 7, 10, 13, 16, 20, 25, 30],
    'min_samples_split': [2, 5, 10],
    'min_samples_leaf': [1, 2, 4, 6]
}

rf_grid_search = GridSearchCV(
    estimator=rf_model,
    param_grid=rf_param_grid,
    scoring='f1_macro', #'roc_auc',
    cv=skf,
    verbose=1

)
```

Best Parameters for Random Forest:

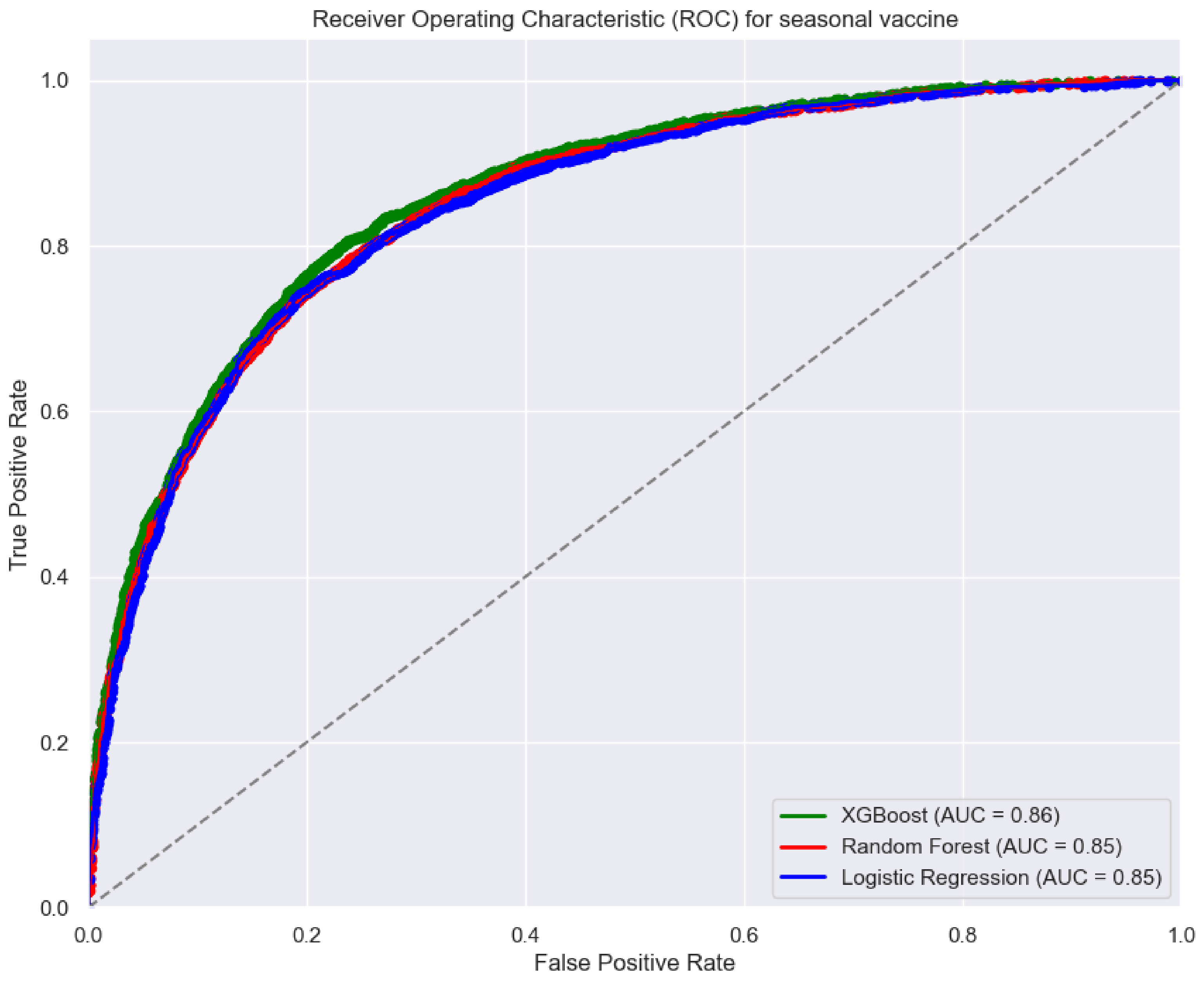
`max_depth=25, min_samples_leaf=2, n_estimators=200, min_samples_split=2`

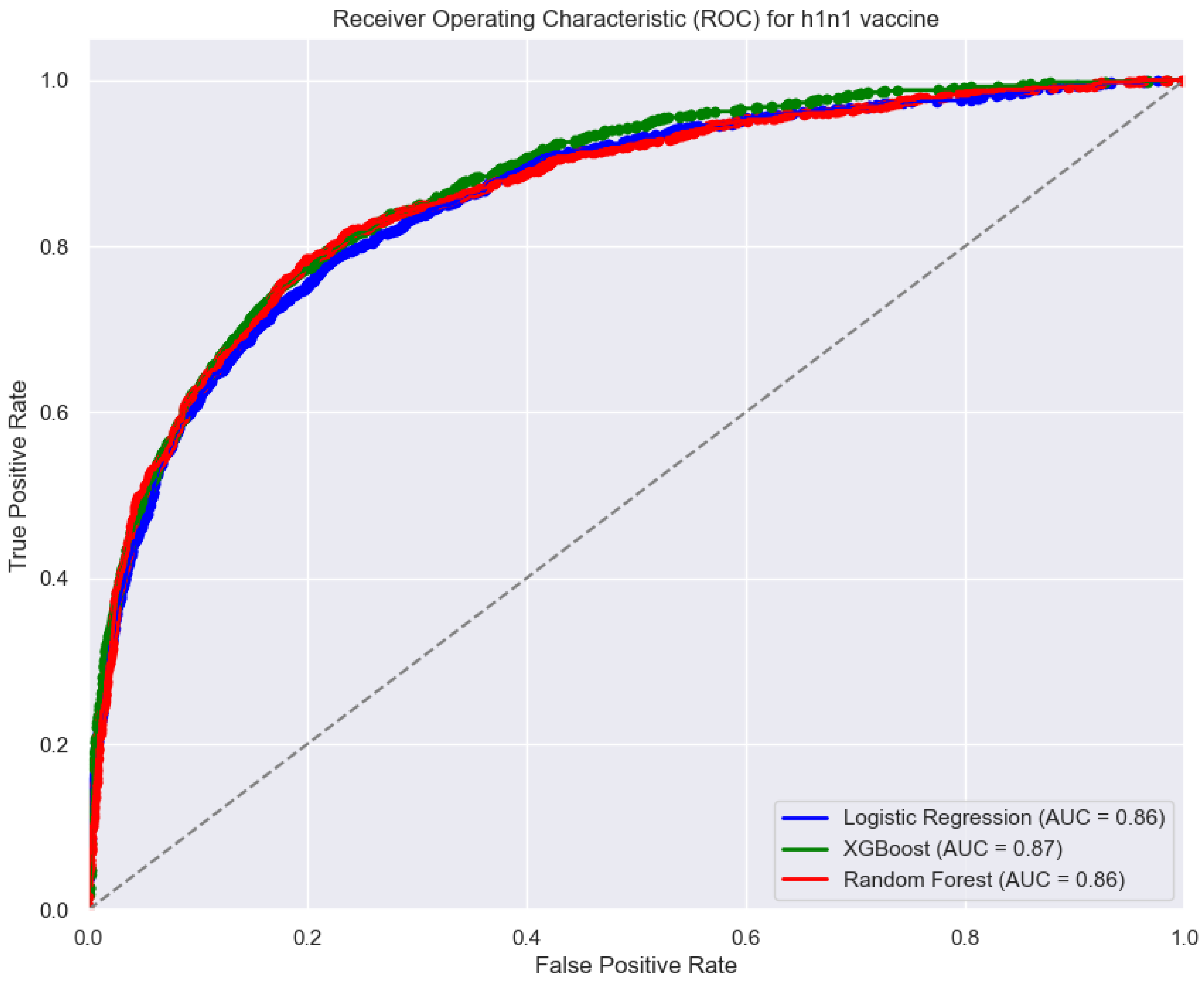
```
xgb_param_grid = {
    'n_estimators': [50, 75, 100, 125, 150, 175, 200, 225],
    'max_depth': [3, 5, 7, 10, 13, 16, 20, 25, 30],
    'min_child_weight': [1, 2, 4],
    'learning_rate': [0.01, 0.1, 0.2]
}
xgb_grid_search = GridSearchCV(
    estimator=xgb_model,
    param_grid=xgb_param_grid,
    scoring='f1_macro',
    cv=skf,
    verbose=1
)
```

Best Parameters for XGBoost:

n_estimators = 225, max_depth = 5, learning_rate = 0.1, min_child_weight=2

MODEL PERFORMANCE EVALUATION





Metric	Regression	Random Forest	XGBoost
ROC-AUC-Score	0.86/0.86	0.86/0.86	0.87/0.86
F1-Score	0.74/0.77	0.75/0.78	0.75/0.78

Skoltech

CONCLUSIONS

Main Conclusions

- *Key Patterns Identified:* The research analysis revealed important patterns that impacted data preprocessing and the choice of machine learning algorithms.
- *Model Used:* The XGBoost model showed better results in terms of ROC-AUC and F1-Score metrics compared to the regression model, which indicates its greater effectiveness in this task.
- *Model Parameters:* The XGBoost model had high n_estimators (around 200+) and low maximum tree depths (up to 5), enhancing its ability to capture complex patterns while preventing overfitting and increasing resilience to noise.
- *Logistic Regression:* Interestingly, the logistic regression model achieved results comparable to XGBoost and Random Forest.

Significance for Public Health

- The analysis results can inform strategies to increase vaccination rates against H1N1 and seasonal influenza, considering population attitudes and behaviors.

Directions for Further Study

- Future research should explore other complex models and ensemble techniques to further enhance predictive capabilities.

Skoltech

THX.

Skoltech