

Лабораторная работа

Методы классификации

Работу подготовили:

Панов Олег, Михаил Боровик, Денис Чашин, Анатолий Мезенов, Никита Бабушкин

Входные данные

- Дана база из 1353 сайтов.
- Сайт может являться фишинговым (702), легитимным (548) или подозрительным (103)
- У каждого сайта есть 9 атрибутов принимающих значения на отрезке $[-1,0,1]$

Атрибуты

Атрибуты имеют следующие названия:

- Server Form Handler
- popUpWindow
- SSLfinal_State
- Request_URL
- URL_of_Anchor
- web_traffic
- URL_Length
- age_of_domain
- having_IP_Address

Постановка задачи

В качестве задания требуется **классифицировать** сайты по заданным параметрам на фишинговые, подозрительные и легитимные. Классифицировать сайты будем с использованием различных **методов классификации**.

Полученные результаты всех методов **сравним** между собой.

Грубо говоря, натренировать модели классификации данных, выбрать **лучшую** из худших полученных моделей и показать полученные результаты.

Выбор моделей

Для решения проблемы классификации мы решили выбрать следующие модели, и распределили их между собой.

- **LDA** - Михаил Бабушкин
- **GNB** - Денис Чашин
- **DT** - Никита Боровик
- **KNN** - Олег Панов
- **SVC** - Анатолий Мезенов

И мы расскажем о них подробнее, но сначала

Что такое GridSearch? Почему мы его используем?

GridSearchCV – это очень мощный инструмент для автоматического подбора параметров для моделей машинного обучения. Метод **поиска по сетке** находит наилучшую комбинацию параметров, которые дают **наименьшую ошибку**, путем обычного перебора: **он создает модель для каждой возможной комбинации параметров**.

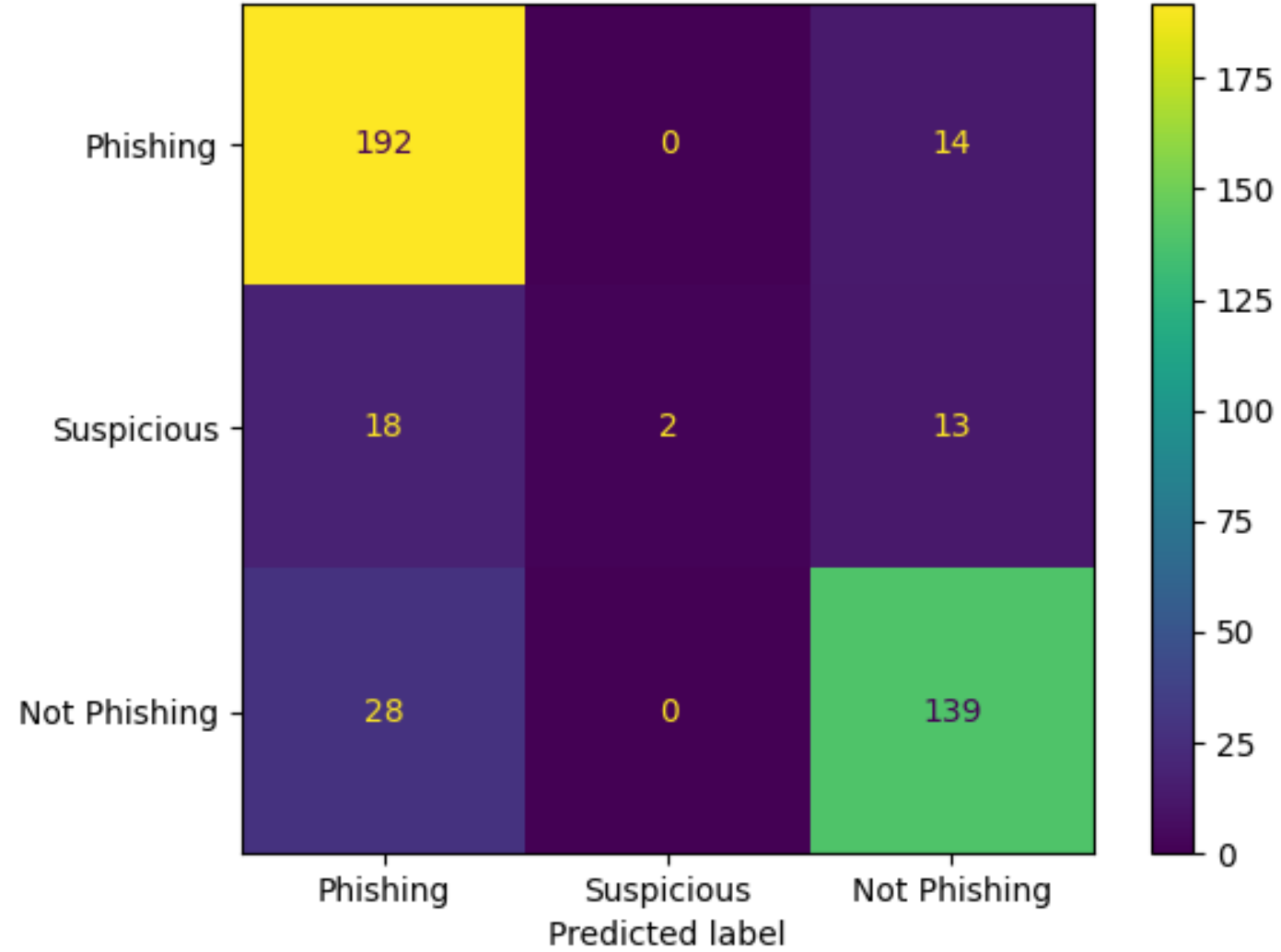
Ну а теперь перейдем к моделям

Модель LDA

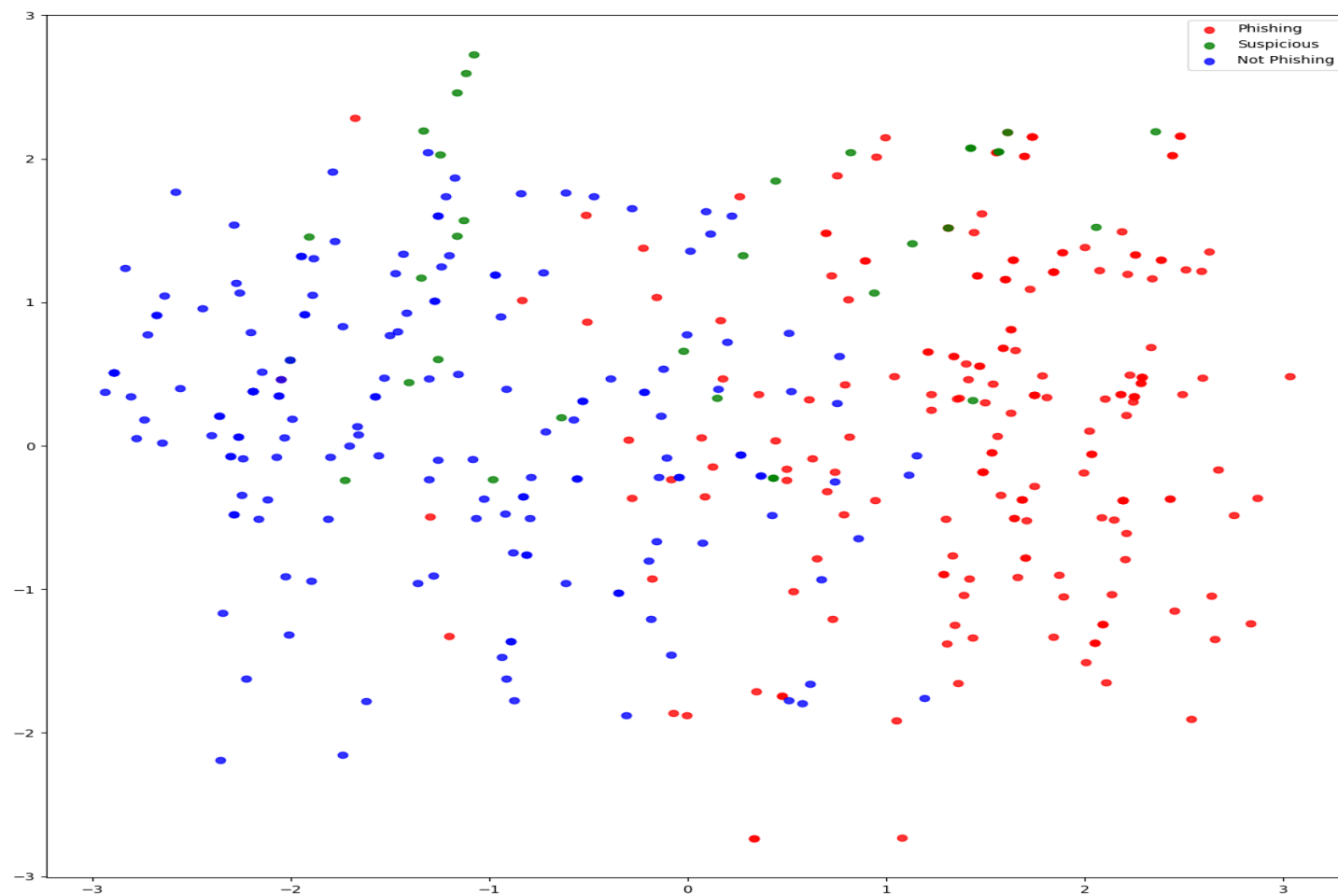
Linear Discriminant Analysis (Линейный Дискриминантный Анализ)

Модель **LDA** представляет собой **метод снижения размерности**, который находит линейные комбинации признаков, максимизирующие разделение между классами. Он стремится максимизировать отношение разброса **между классами** к разбросу **внутри классов**.

- Она **хороша** потому что отлично подходит для задач с **высокой размерностью данных**
- **Но** также она слишком чувствительна к выбросам, а также предполагает нормальное распределение данных



Matrix for LDA model



LDA visualisation

Результат работы модели

- Обучающая выборка: точность в 82.0%
- Тестовая выборка: точность в 82.02%
- Время работы: 64.21 секунд

Подобранные параметры

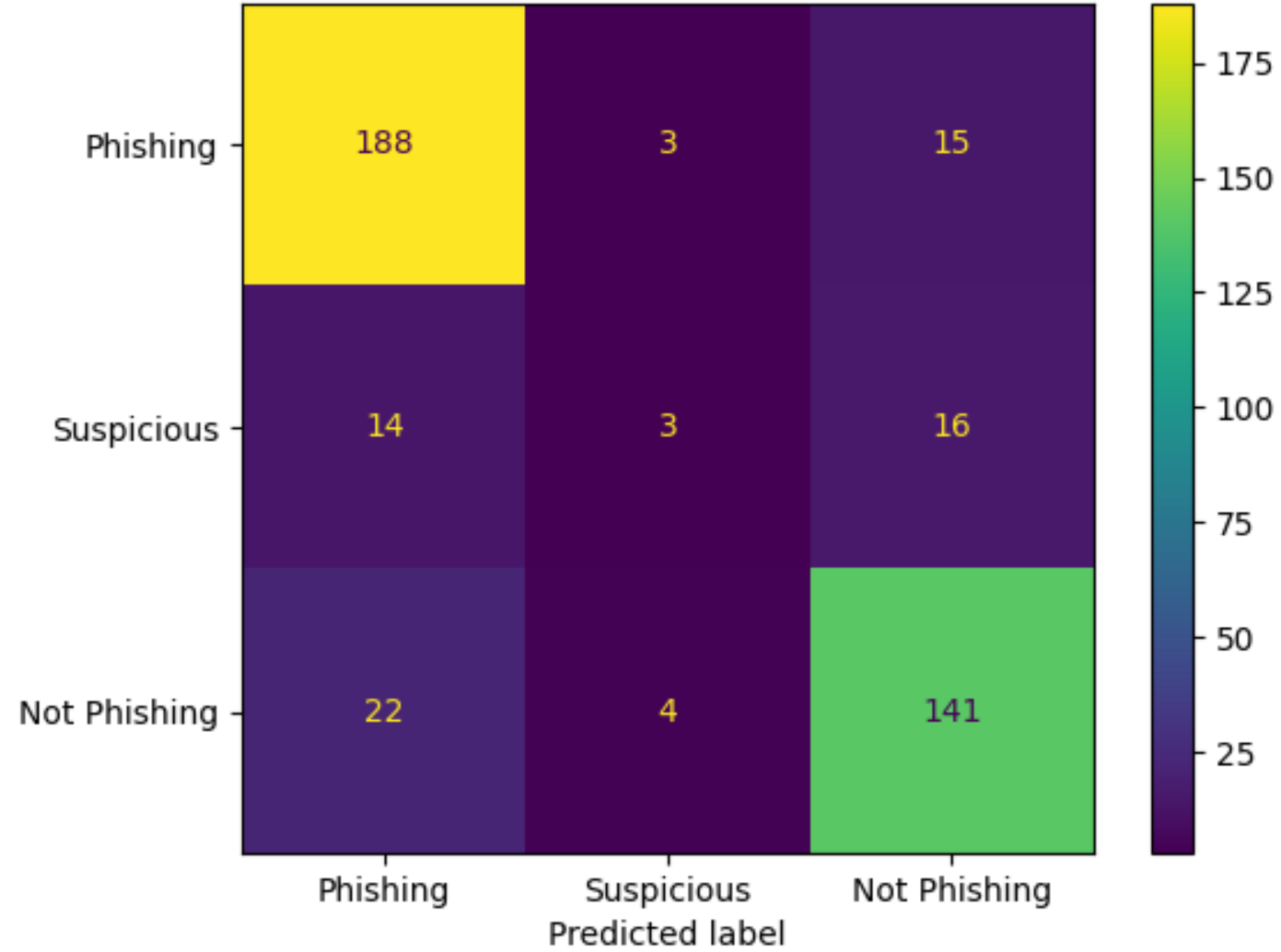
- Выбранный критерий: Eigen
- Shrinkage (*сокращение*) : 0.01

Модель GNB

Gaussian Naive Baías (Наивный Байесовский Классификатор)

Модель **GNB** представляет собой метод классификации, который предполагает, что **признаки независимы внутри каждого класса**.

- Она **удобна** в том, что использует нормальное распределение для оценки вероятностей, что делает его **простым**.
- **Но есть минусы**, ведь она не учитывает возможные взаимосвязи между признаками.



Matrix for GNB model

Результат работы модели

- Обучающая выборка: точность в 81.35%
- Тестовая выборка: точность в 81.77%
- Время работы: 1.30 секунд

Подобранные параметры

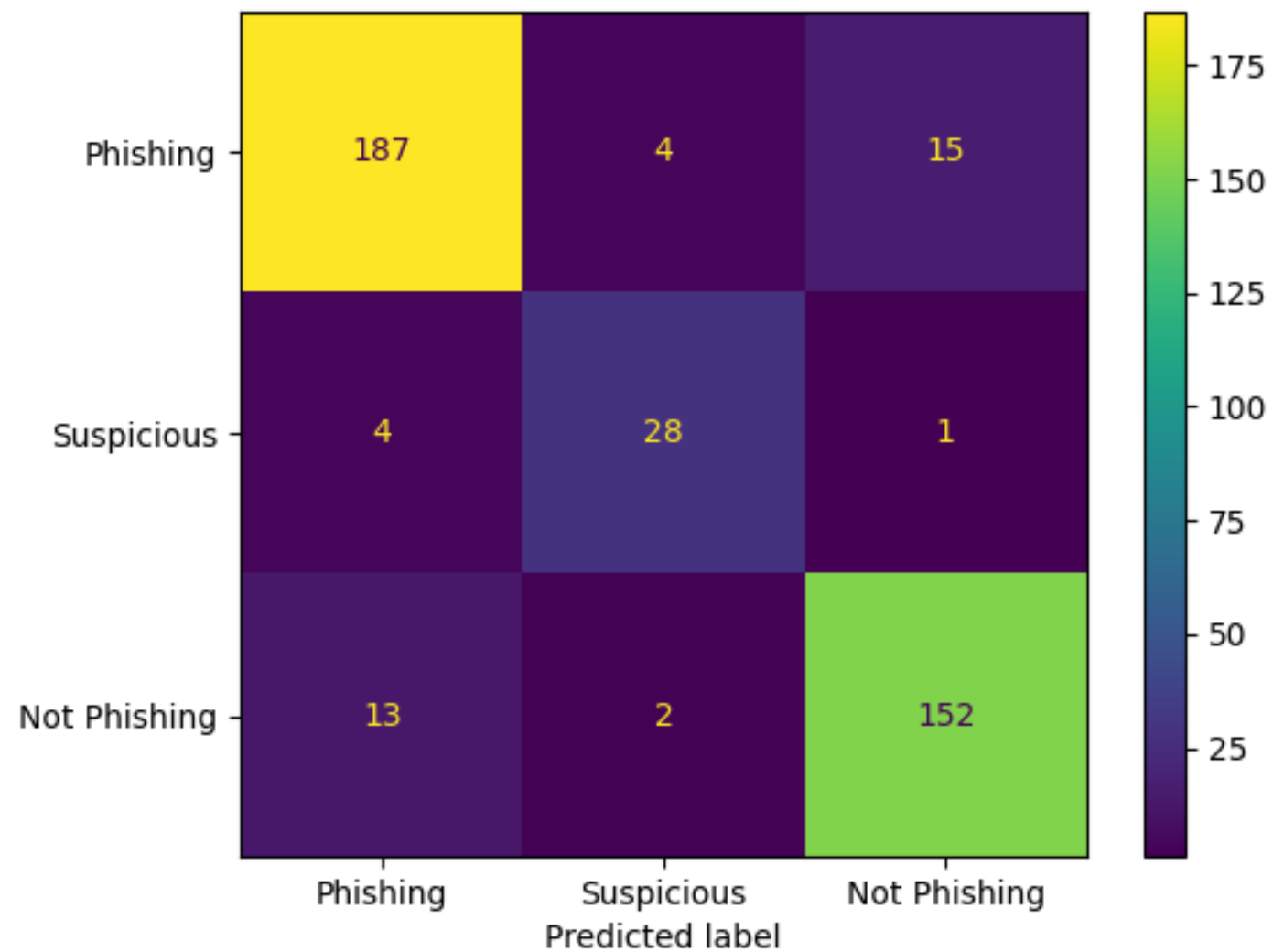
- priors : None
- var smoothing : 10^{-9}

Модель DTC

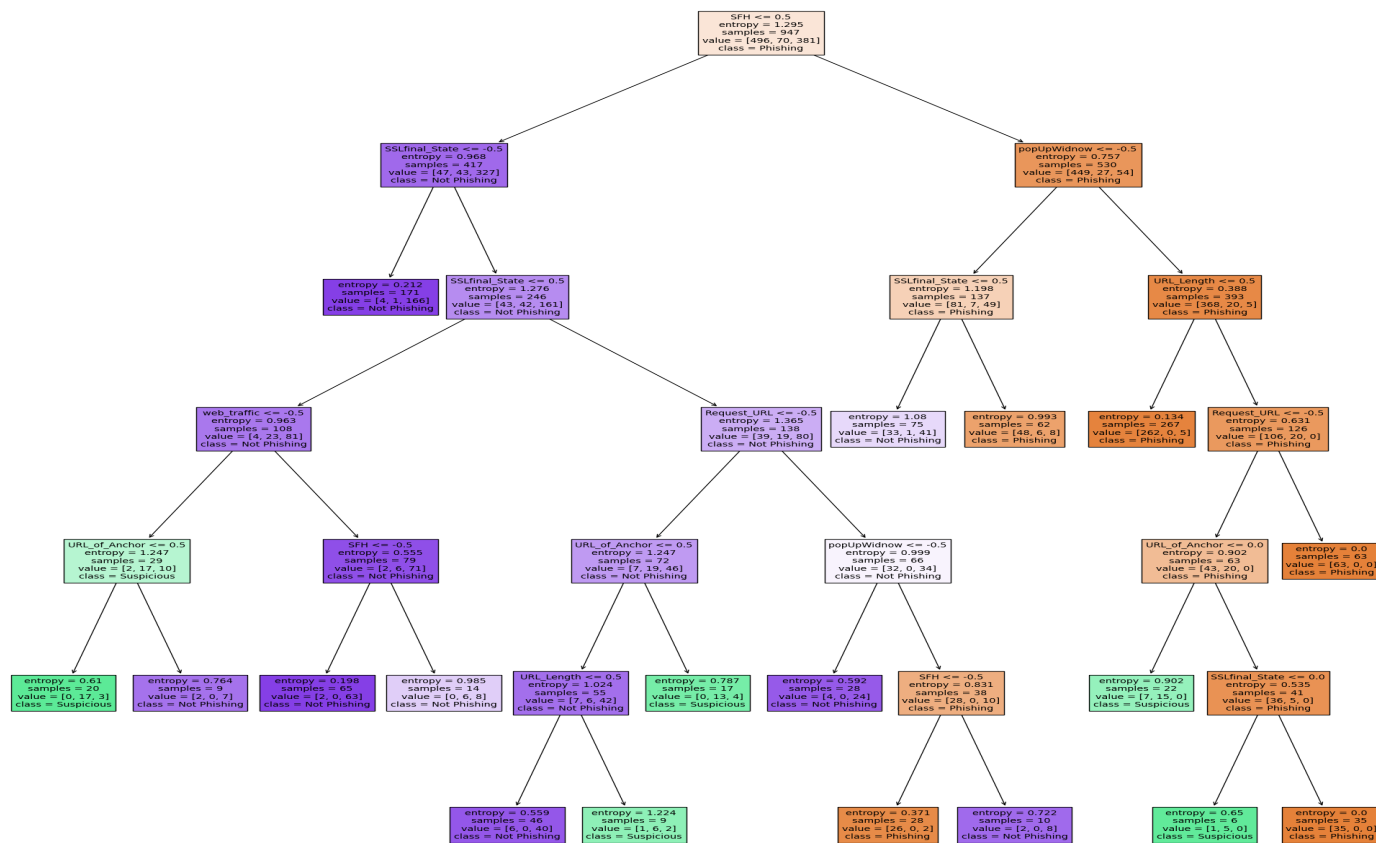
Decision Tree (Классификатор Деревя Решений)

DT - это метод классификации, основанный на построении **дерева решений**, где каждый узел представляет собой тест по одному из признаков, а листья соответствуют классам.

- Чем глубже дерево, тем сложнее правила принятия решений и тем лучше модель.
- Она **хорошо** интерпретируема и способна обрабатывать нелинейные зависимости в данных.
- **Однако** могут создаваться слишком сложные деревья, которые плохо обобщают данные.



Matrix for DT model



Desision Tree visualisation

Результат работы модели

- Обучающая выборка: точность в **87.33%**
- Тестовая выборка: точность в **90.39%**
- Время работы: **168.73** секунд

Подобранные параметры

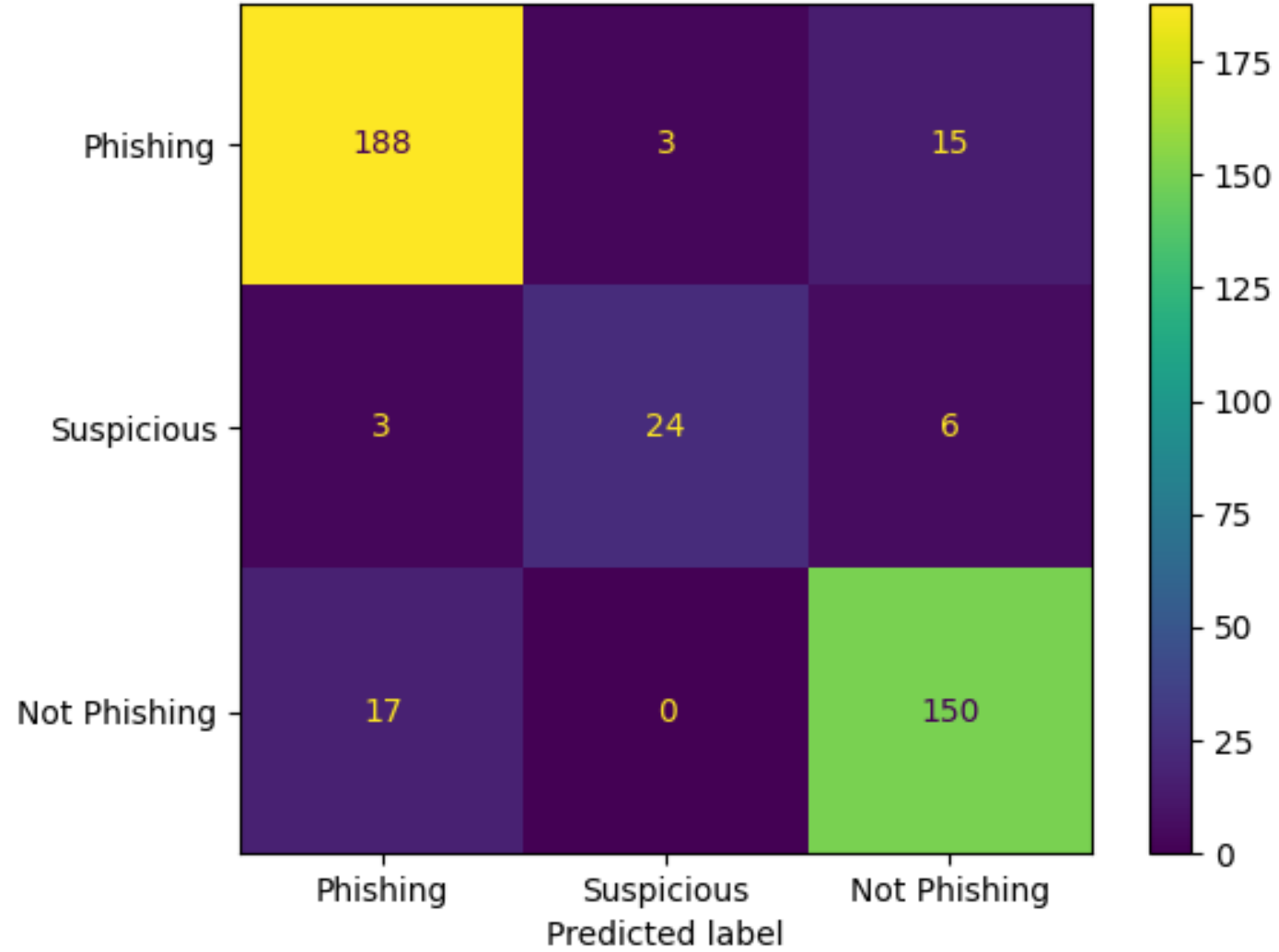
- Выбранный критерий: **entropy**
- Глубина дерева: **6**
- Максимальное к-во листьев: **18**
- Минимальное к-во экземпляров в листе: **2**

Модель KNN

K Nearest Neighbors (Классификатор "k-ближайших соседей")

Модель KNN основана на принципе отнесения объекта к классу, к которому принадлежат его **ближайшие соседи** по признакам.

- Модель **хороша**, поскольку ее метод не использует сложную математику, а реализация проста и очевидна.
- **Однако** она может быть вычислительно затратной для больших наборов данных.



Matrix for KNN model

Результат работы модели

- Обучающая выборка: точность в 87.23%
- Тестовая выборка: точность в 89.16%
- Время работы: 45.11 секунд

Подобранные параметры

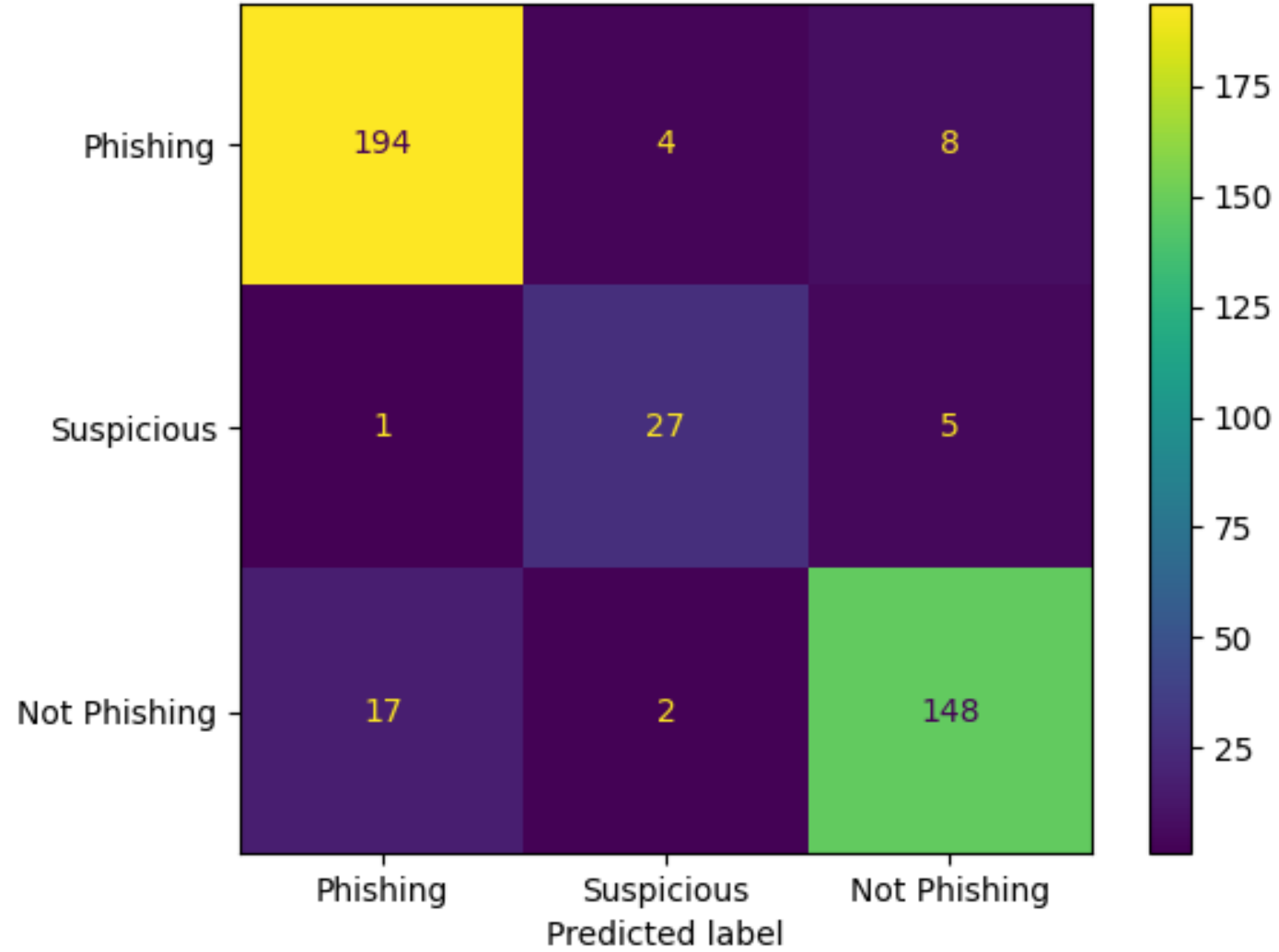
- Выбранная метрика: Минковский ($p = 1$)
- К-во листов: 7
- К-во соседей: 5
- Параметр веса: Единый

Модель SVC

Support Vector Classification (Метод Опорных Векторов)

Модель SVC строит разделяющую гиперплоскость между классами, оптимизируя расстояние до ближайших точек каждого класса.

- Эта модель **хорошо** работает в задачах с нелинейными зависимостями и позволяет достичь высокой точности классификации
- **Однако** требует внимательной настройки параметров для достижения оптимальных результатов.



Matrix for SVC model

Результат работы модели

- Обучающая выборка: точность в **94.89%**
- Тестовая выборка: точность в **94.62%**
- Время работы: **53.94** секунд

Подобранные параметры

- Выбранный критерий kernel: **rbf**
- C: **10**
- Гамма: **1**

Итоги

В ходе проделанной работы было выявлено, что наилучшей моделью для классификации фишинговых сайтов оказалась SVC (Метод Опорных Векторов) с точностью на тестовой выборке в более чем **94%!!!** 🎉🎉🎉

Почему так?

- в данных не выявлены явные линейные зависимости
- экспериментальная балансировка с дополнительным обучением
- так вышло 😊😊😊

Список Литературы

- Документации Sklearn, Pandas, Numpy
- Статьи с habr.com, [Tproger.com](https://tproger.com)
- UCI Machine Learning Repository
- Лекции с курса "Приложения и практика анализа данных"

Старались,

Data Balbesing ❤️

это мы кайфуем от
максимальных баллов за лабу

->

