Лабораторная работа

Методы классификации

Работу подготовили:

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

Входные данные

- Дана база из **1353** сайтов.
- Сайт может является фишинговым *(702)*, легитимным *(548)* или подозрительным *(103)*
- У Каждого сайта есть **9** аттрибутов *принимающих значения **[-1,0,1]***
- 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 тренированных и показать лучшие 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?

> и почему мы его используем

GridSearchCV – это очень мощный инструмент для автоматического подбирания 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**.

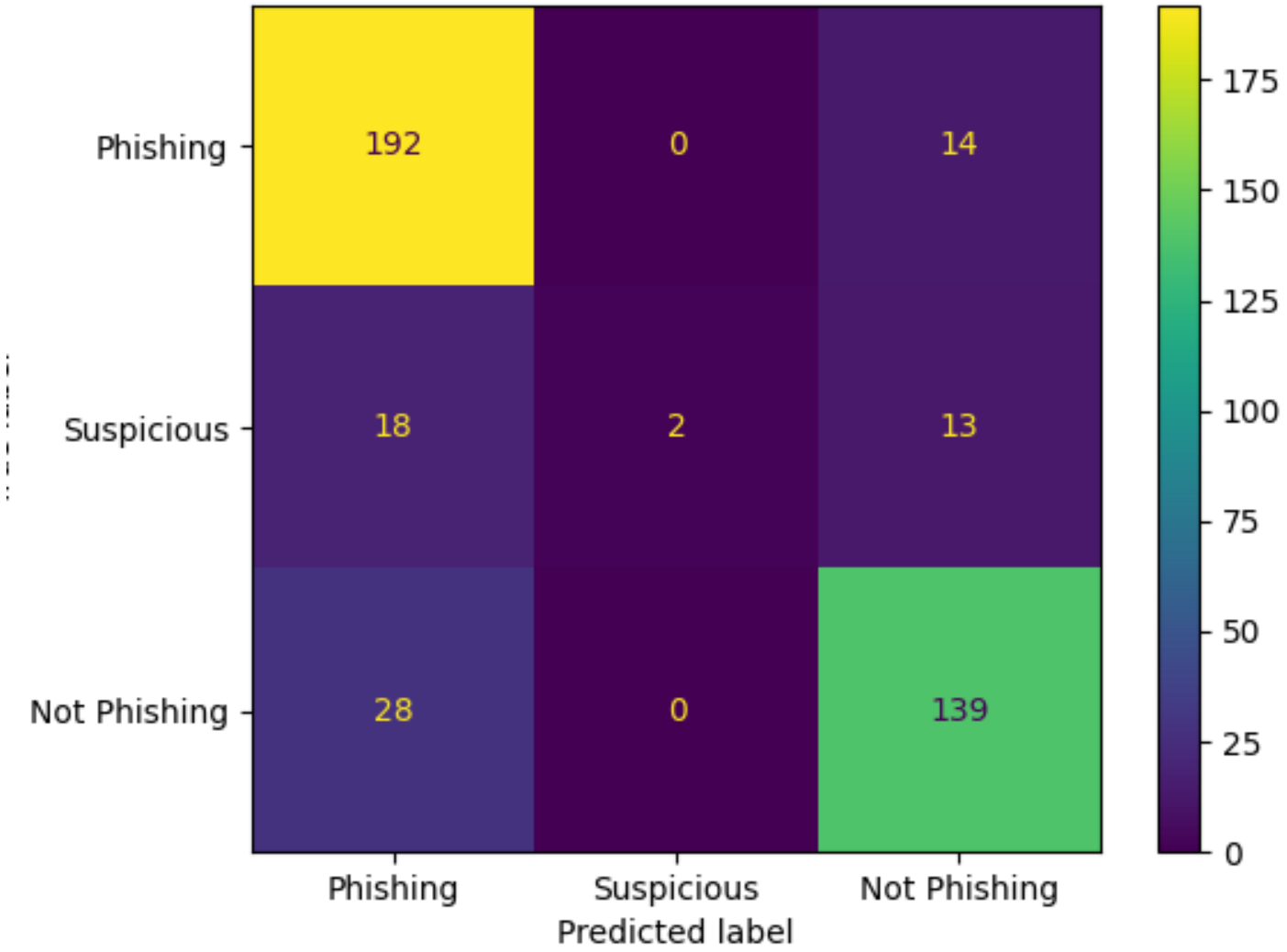
> Ну а теперь перейдем к моделям

Модель LDA

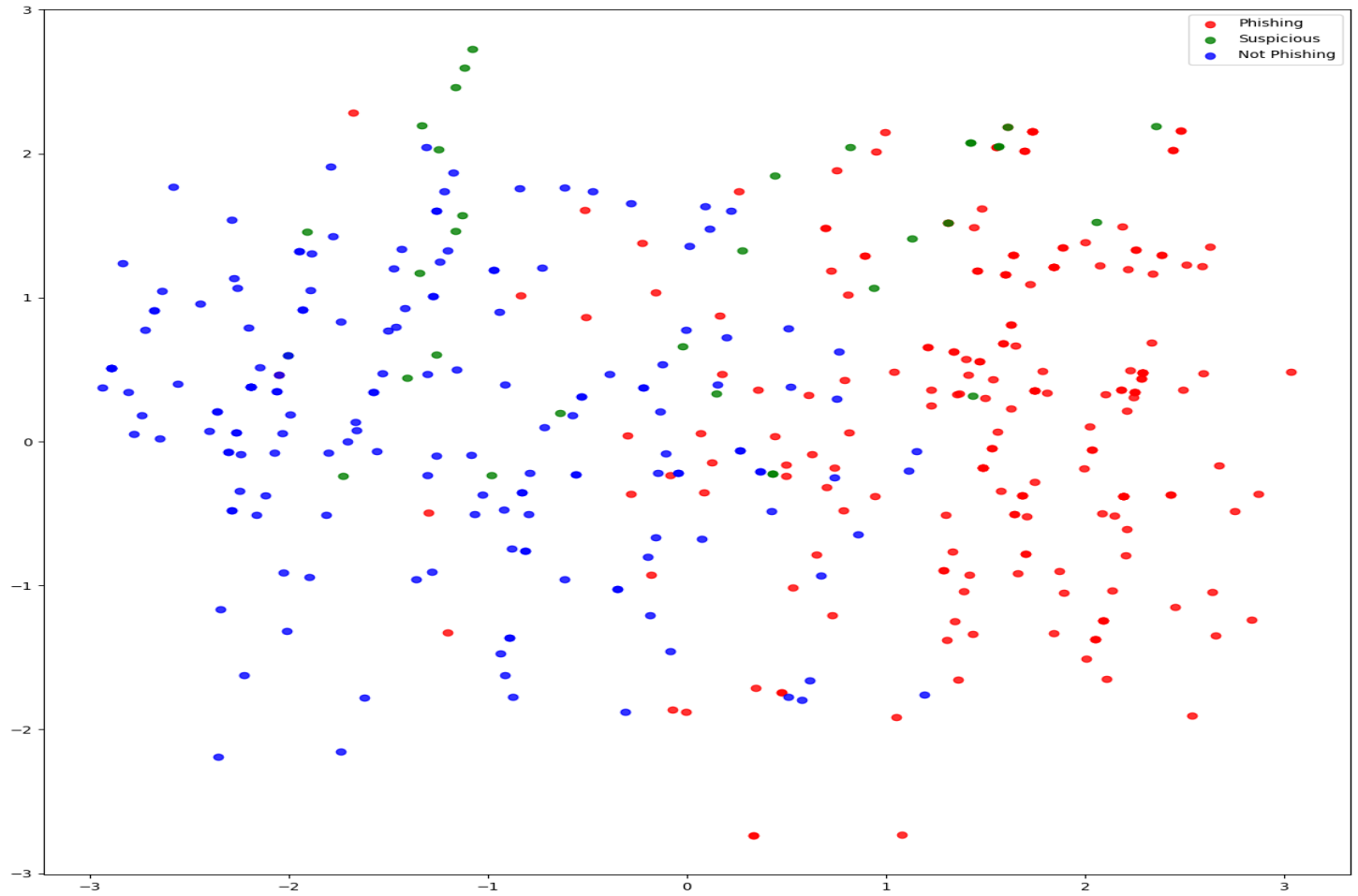
> Linear Discriminant Analysis (линейный дискриминантный анализ)

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 преполагает 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

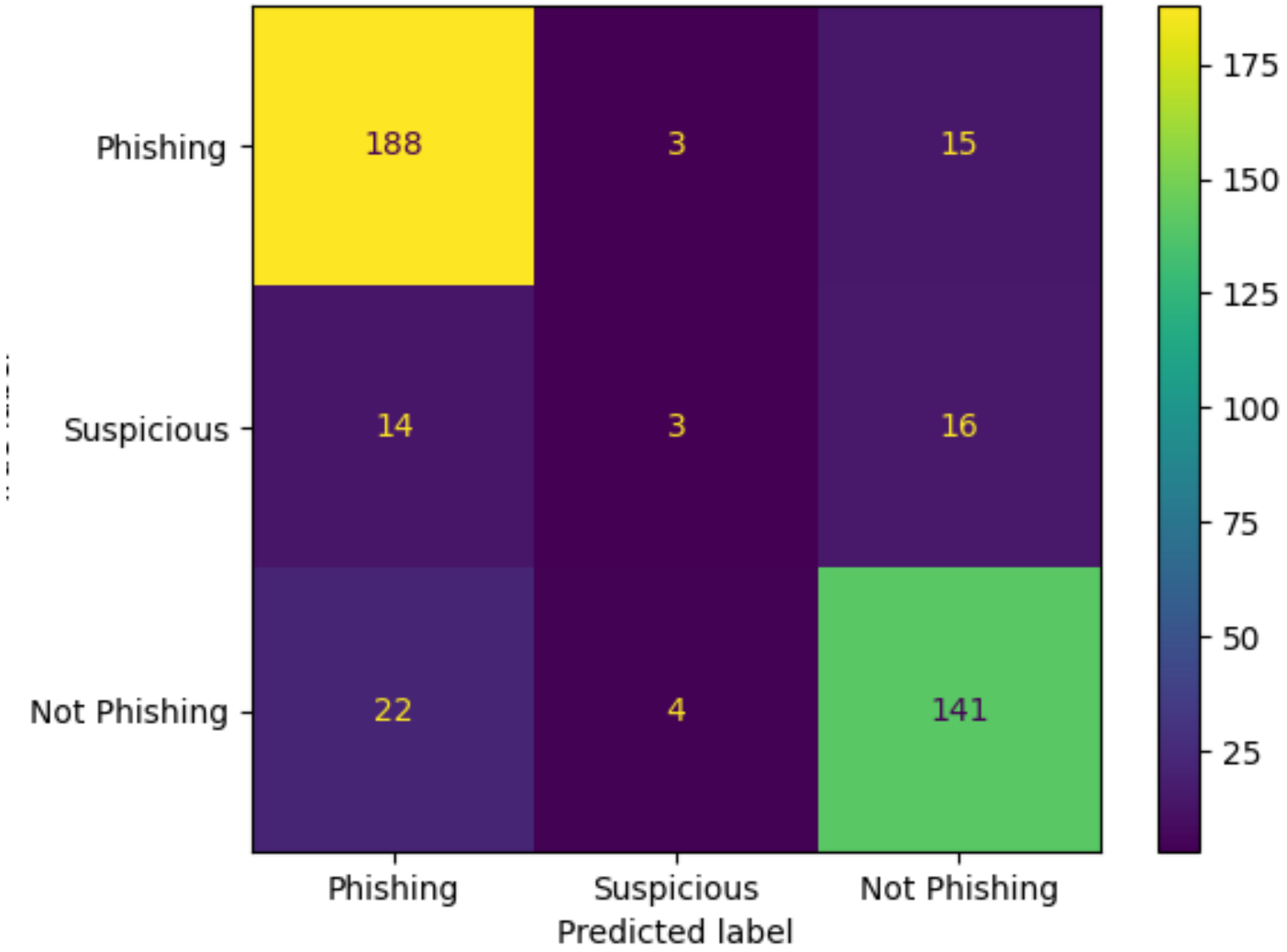
- Выбранный критерий: **Eigen**
- Shrinkage *(сокращение)* : **0.01**

Модель GNB

> Gausian Naive Baias (наивный байесовский классификатор)

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

- Она **удобна** в в том, что использует нормальное распределение для оценки вероятностей, что делает его **простым**
- **Но есть беда**, ведь он 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

- var smoothing : **10^-9**

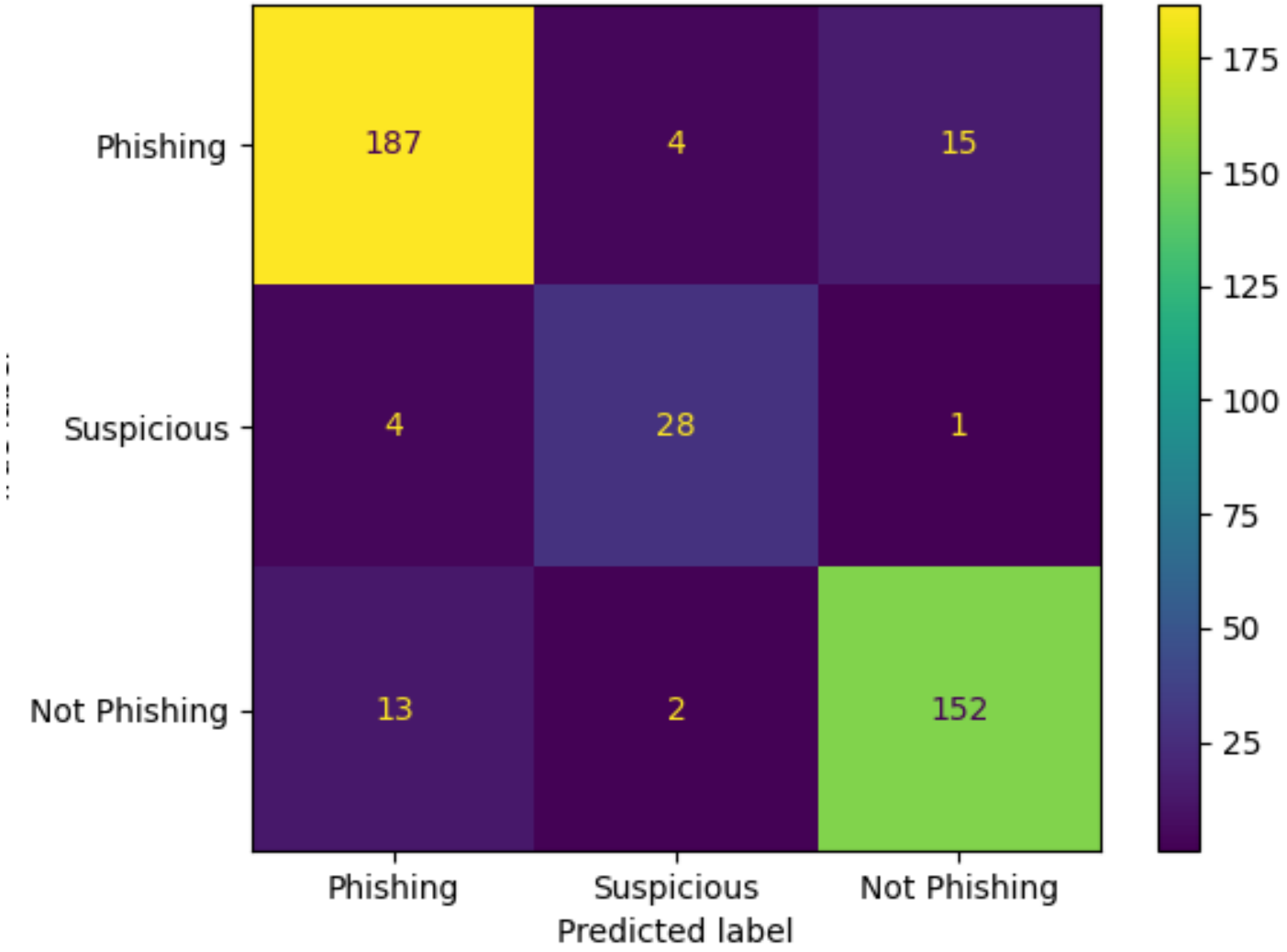
Модель DTC

> Decision Tree (Классификатор дерева решений)

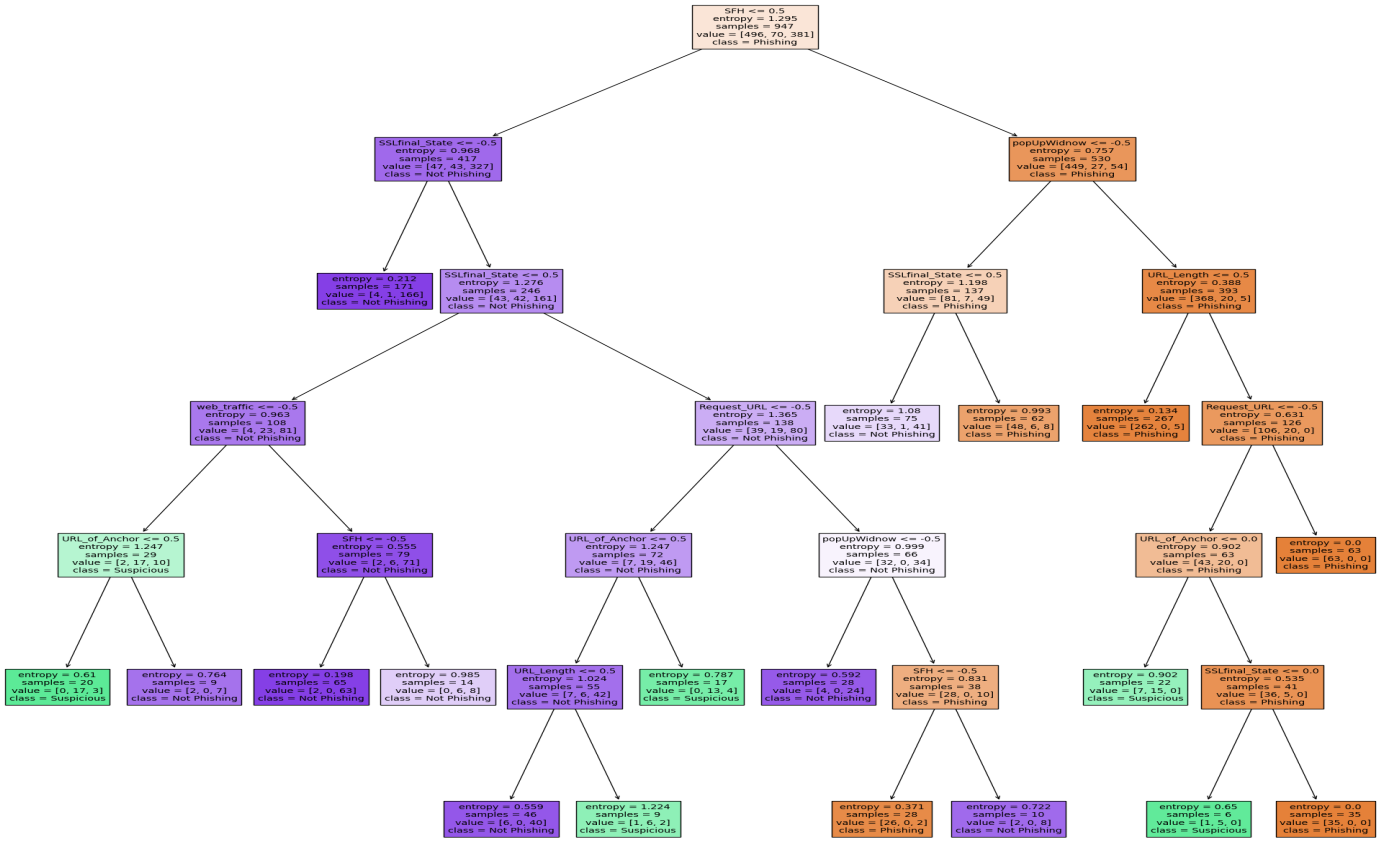
DT - это метод классификации, основанный на построении **дерева решений**, где каждый узел представляет собой тест по одному из признаков, а листья соответствуют классам

> Чем глубже дерево, тем сложнее правила принятия решений и тем лучше модель

- Она хорошо интерпретируема и способена обрабатывать нелинейные зависимости в данных
- Но, 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

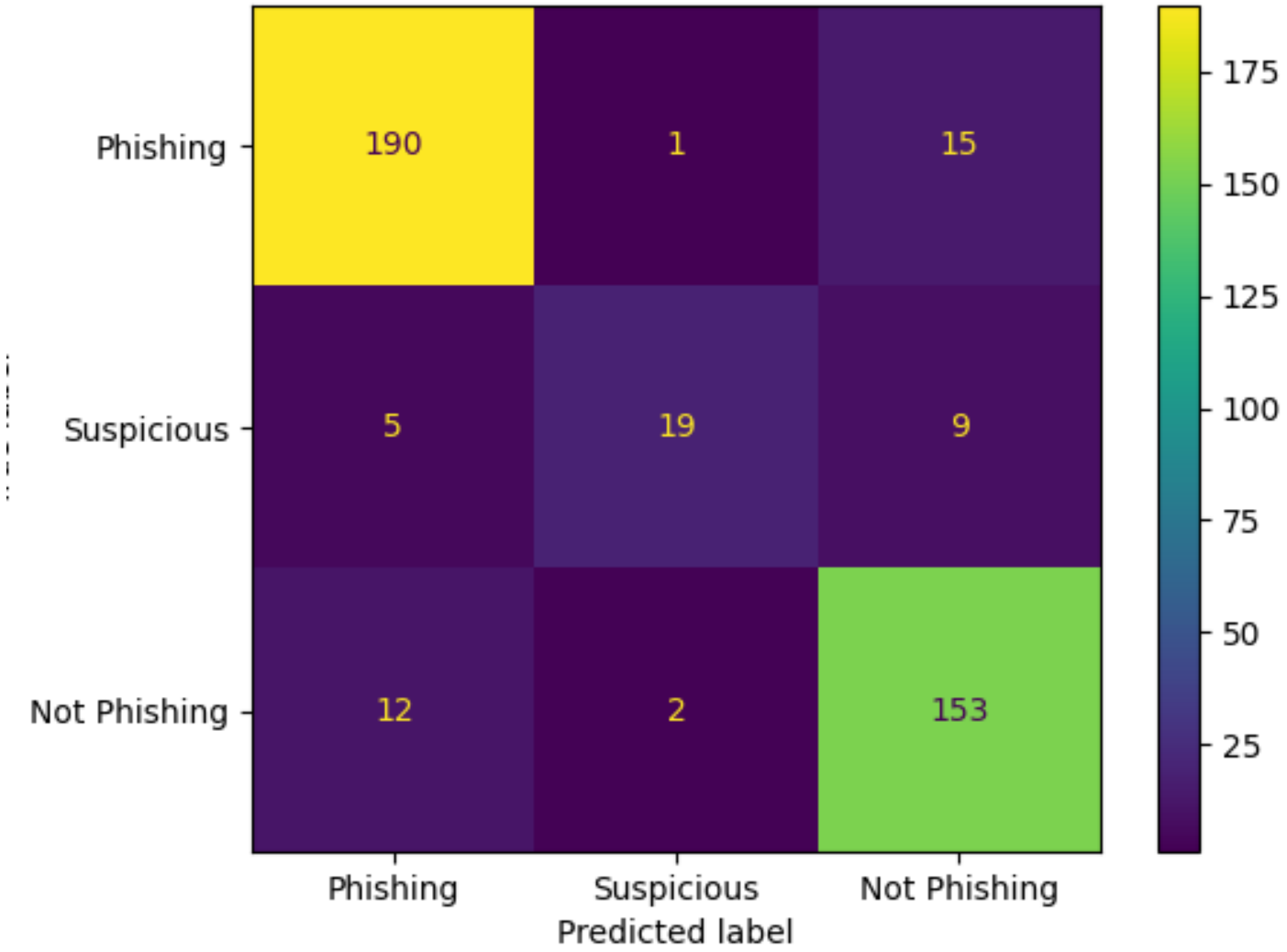
- Выбранный критерий: **entropy**
- Глубина дерева: **6**
- Максимальное к-во листьев: **18**
- Минимальное к-во экземпляров в листе: **2**

Модель KNN

> K Nearest Neighbors (классификатор "k-ближайших соседей")

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

- Модель хороша, тк ее метод не использует сложную математику и реализация проста
- 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

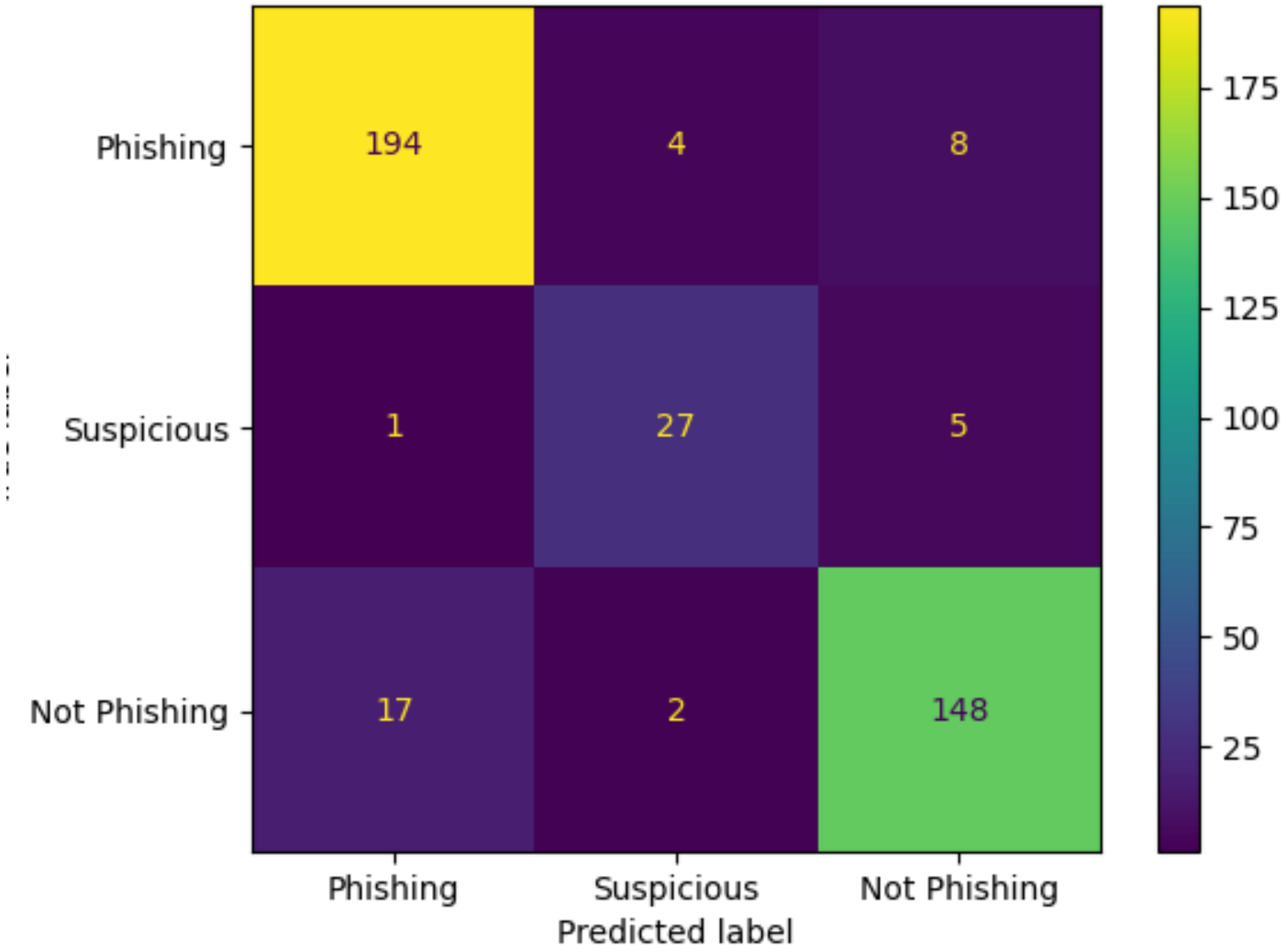
- Выбранная метрика: **Минковский** *(p = 1)*
- К-во листов: **7**
- К-во соседей: **5**
- Параметр веса: **Единый**

Моедль SVC

> Support Vector Classification (метод опорных векторов)

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

- Этот модель хорошо работает в задачах с нелинейными зависимостями и позволяет достичь высокой точности классификации
- но 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 качестве итога мы определили, что лучшей моделью с оптимальными параметрами оказалась *(ОНА)*, вот ЕЕ результаты:

-
-

Старались,

Data Balbesing ❤️

> Никита смотри какой там негр идет!!!! ■_■