Алгоритмы машинного обучения для

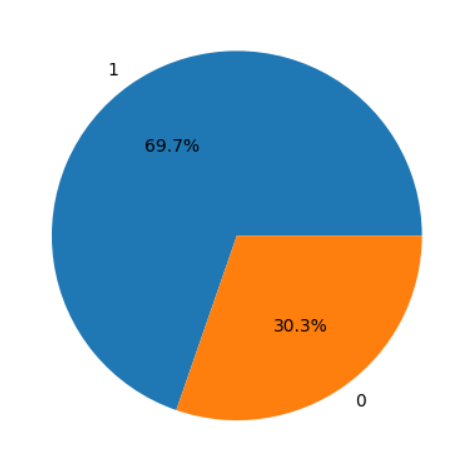
предсказания наличия или отсутствия

нестандартных вихревых явлений

1. Подготовка данных

Оснавная идея в том, чтобы найти взаимасвязи между данными, разметить данные по лайблам(они же метки), а так же подготовить их для того, чтобы потом использовать алгоритмы ml.

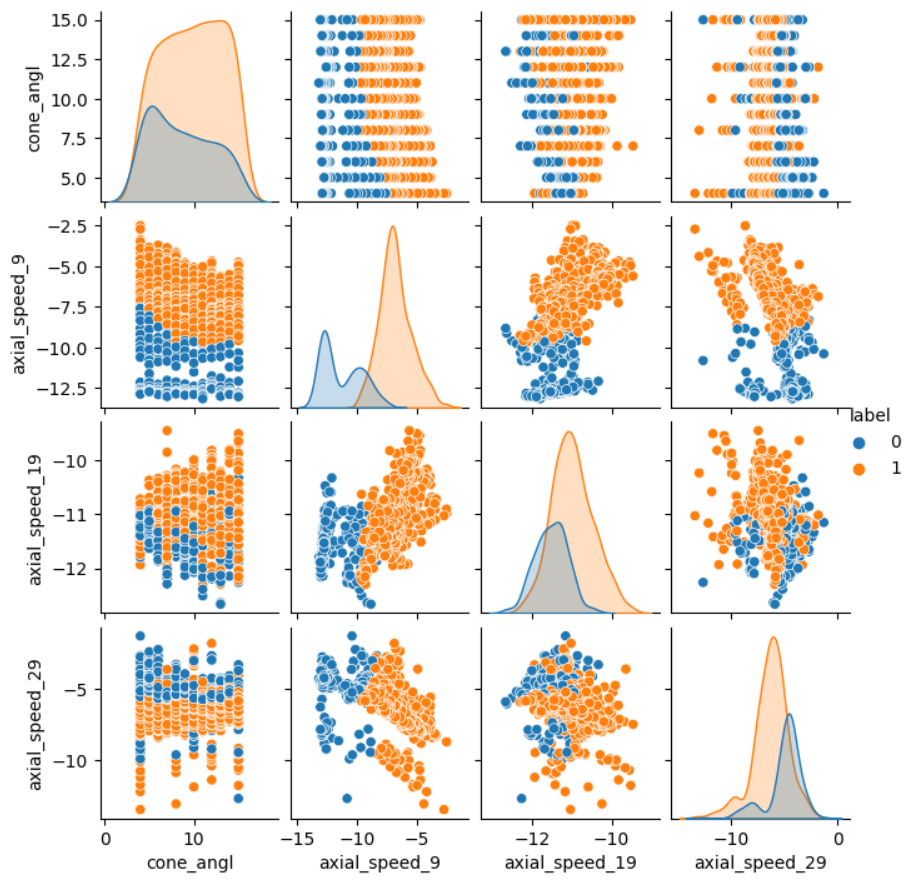
Заметим, что явно виден дизбаланс между классами



1 - вихревые явления есть

0 - вихревых явлений нету

Так же использовалась библиотека для python “seaborn”, которая показывает, как распределенны классы по двум признакам.



Здесь представленна часть этой таблицы, наглядно видно, что параметр cone\_angl в сочетание с некоторыми параметрами, может явно разделить классы.

1. Работа над моделью ml

Так как классы распределенны неравномерно, то тренеровочные и тестовые выборки подбирались с помощью библиотеки для python “train\_test\_val”, которая разделяет даннык равномерно.

В итоге самые лучшие метрики показала модель RandomForestClassifire из библиотеки “sklearn”.

Основные метрики - accuracy, recall, precision и f1

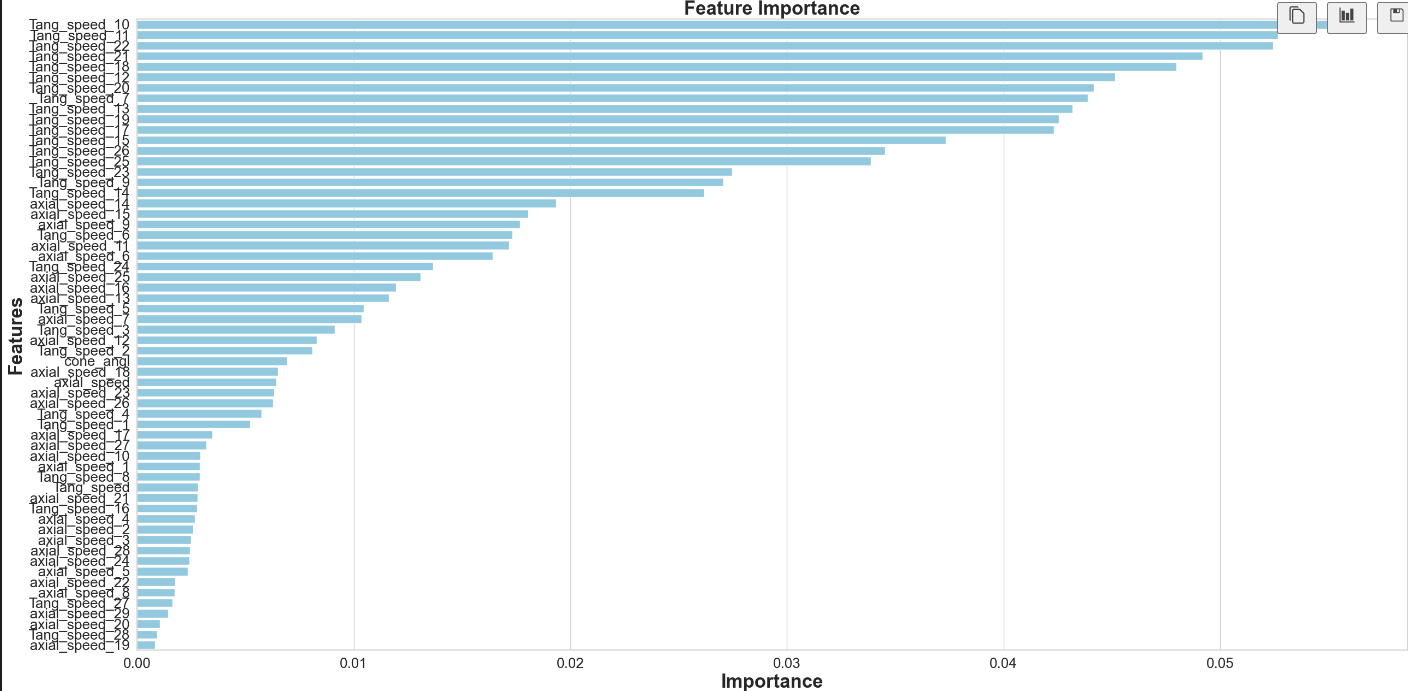
Метрики, кроме accuracy, используются т.к в выборке будет присутствовать явный дизбаланс классов.

Accuracy = 0.961

Precision = 0.963

Recall = 0.971

Кроме того, RandomForest показывает, насколько модель часто использует определённые параметры.



Видно, что cone\_angl, не является одним из самых важных метрики. Т.к. этот параметр, по идее, должен являтся одним из самых важных, можно сделать выводы, что это негативно сказывается на метрике и итоговой работе модели.

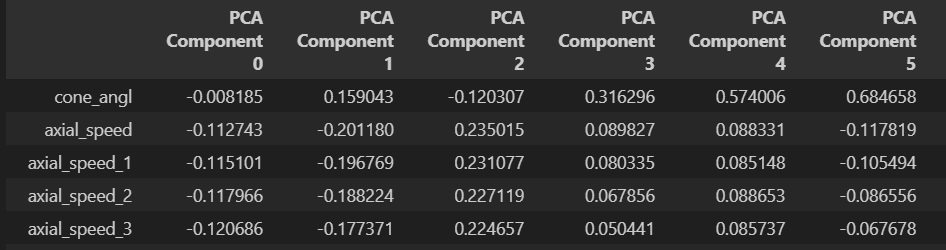
1. Доработка модели

Первая идея - понижать количество параметров, удаляя параметры, котроые не являются важными для нашей модели и посмотреть, повысится ли важность параметра cone\_angl и в частности метрика.

Но в этом случае параметр cone\_angl не остался среди пяти самых важных параметров.

Альтернативная возможность, это используя “Метод главных компонент”, он же “principal component analys”.

Он позволяет уменьшить размерность данных, не теряя при этом самую важную информацию.



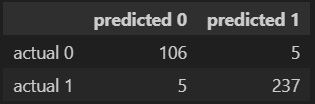
Суть метода в вычисление ковариационной матрицы, для переданных данных. В ячейке у таблицы содержиться результат ковариации для двух параметров, по свойству ковариации на главной диаганале будет лежать дисперсия от одного параметра.

Формально задача состоит в нахождение пространства данных, для которых разброс данных максимален (то есть максимально среднеквадратичное отклонение от среднего).

В примере, который представлен выше, PCA componetn 0, 1, 2 и.т.д обозначают соответствующие данные из нашей выборки.

Чем больше значение ячейки - тем больше эта пара параметров влияет на нашу модель. Поэтому задача состоит в отборе самых важных параметров, по этой таблице.

После попытки использовать модель с указанными параметрами повысилась и метрика и важность параметра cone\_angl.



Итоговые метрики на этом этапе

Accuracy = 0.987

Precision = 0.991

Recal = 0.971

Но в этом случае - метрика могла повысится за счёт отбора наиболее важных параметров, а в тоже время, упасть из-за недостаточного количества параметров при обучение модели. В итоговой моделе были добавленны ещё два параметра, но возможно можно дальше добавлять и отбирать наиболее задействуемые параметры, согласно таблице, составленной с помощью Метода главных компонент.