1. Сформулируйте задачу классификации?

задано конечное множество объектов, для которых известно, к каким классам они относятся. Это множество называется обучающей выборкой.

1. Что означает обучение с учителем?

данные в обучающей выборке должны быть размечены или каждому набору признаков должны соответствовать метки

1. Зачем разделять обучающую выборку?

Это позволяет проверить, как модель обобщает знания на новых данных. Если данные отсортированы по какому-либо признаку, разделение без перемешивания может привести к ошибкам, поэтому данные предварительно перемешивают

1. Что означает переобученная модель? Как с этим бороться?

Переобученная модель слишком точно описывает обучающую выборку, включая шум и выбросы, но плохо обобщает на новые данные.

Увеличить объем обучающих данных.

* Упростить модель
* Использовать регуляризацию
* Применять ансамблевые методы, такие как случайный лес.

1. Что означает обобщающая способность моделей машинного обучения?

это её способность корректно предсказывать результаты на новых, ранее не виденных данных.

1. Объясните значения в матрице ошибок, как она рассчитывается?

это таблица, сравнивающая предсказания модели с фактическими значениями. Для бинарной классификации она содержит 4 значения:

* **True Positive (TP)**: модель правильно предсказала положительный класс.
* **True Negative (TN)**: модель правильно предсказала отрицательный класс.
* **False Positive (FP)**: модель ошибочно предсказала положительный класс (ложное срабатывание).
* **False Negative (FN)**: модель ошибочно предсказала отрицательный класс (пропуск цели).

1. Что показывают accuracy, precision и recall?

 **Accuracy (точность)**: доля правильных предсказаний модели от общего числа предсказаний, рассчитывается как (TP + TN) / (TP + TN + FP + FN).

 **Precision (точность класса)**: доля правильно предсказанных объектов положительного класса от всех предсказанных как положительные, рассчитывается как TP / (TP + FP).

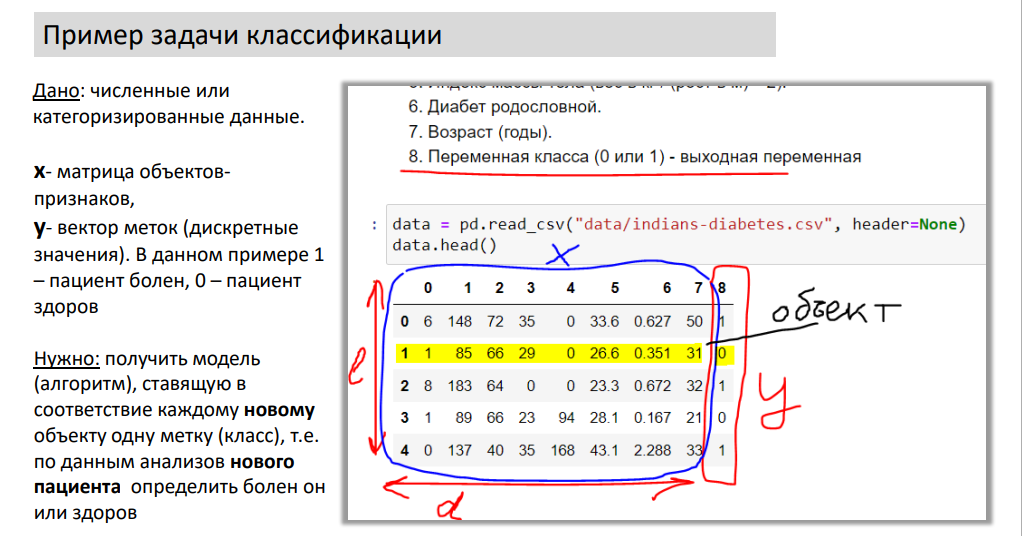
 **Recall (полнота)**: доля правильно предсказанных объектов положительного класса от всех истинных положительных объектов, рассчитывается как TP / (TP + FN).

1. Что означает понятие ансамбль в машинном обучении?

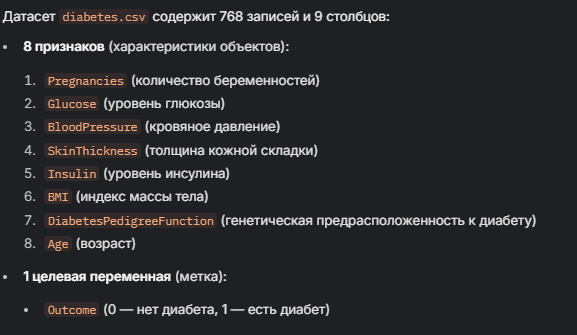
метод, объединяющий несколько моделей для улучшения общей точности и устойчивости предсказаний.

1. Расскажите о методе случайного леса.

это ансамблевый метод, основанный на множестве деревьев решений. Каждое дерево обучается на случайной подвыборке данных и случайном подмножестве признаков. Итоговое предсказание формируется путем голосования или усреднения результатов всех деревьев.



Метод деревьев решений для задачи классификации состоит в том, чтобы осуществлять процесс деления исходных данные на группы, пока не будут получены однородные их множества.



Модель **k-ближайших соседей (k-Nearest Neighbors, KNN)** — это алгоритм машинного обучения, используемый для задач классификации и регрессии. В контексте лабораторной работы она применяется для задачи классификации (предсказание наличия диабета).

GridSearchCV — это инструмент из библиотеки scikit-learn, используемый для **автоматического подбора оптимальных гиперпараметров** модели машинного обучения.

* 1. В чем суть линейных классификаторов? Какие методы относятся к линейным классификаторам?

Линейные классификаторы разделяют признаковое пространство гиперплоскостью на два полупространства, каждое из которых соответствует одному из двух классов. Если выборка разделяется без ошибок, она называется линейно разделимой.  Логистическая регрессия

 Линейный метод опорных векторов

* 1. Что такое регуляризация?

метод добавления ограничений к модели для предотвращения переобучения.

* 1. Объясните назначение L1 и L2-регуляризации?

Упрощает модель, отбирая наиболее важные признаки и обнуляя менее значимые коэффициенты. Предотвращает переобучение, уменьшая величину весовых коэффициентов, но не обнуляет их.

* 1. Объясните влияние параметров С и гамма на регуляризацию модели?

 Низкие значения C: Модель подстраивается под большинство данных, игнорируя выбросы, что улучшает обобщение

 Высокие значения C: Модель старается правильно классифицировать каждую точку, что может привести к переобучению.

 Низкая гамма: Учитываются даже удалённые точки, гиперплоскость более гладкая, что улучшает обобщение.

 Высокая гамма: Учитываются только ближайшие к гиперплоскости точки, что делает модель чувствительной к выбросам и может привести к переобучению.

* 1. Какой метод помогает подобрать лучшую комбинацию параметров?

используется **поиск по сетке (GridSearchCV)**. Он перебирает комбинации параметров, заданные в param\_grid, и оценивает их производительность с помощью кросс-валидации, возвращая наилучшую комбинацию.

* 1. Что такое специфичность и чувствительность модели?
* Доля истинно положительных случаев, правильно классифицированных моделью.
* Доля истинно отрицательных случаев, правильно классифицированных моделью.
  1. Как по ROC-кривой можно сравнить модели. Что означает AUC в ROCкривой?
* ROC-кривая отображает зависимость чувствительности (TPR) от 1-специфичности (FPR).
* Площадь под ROC-кривой, которая количественно оценивает качество модели.