1. Bài toán phân loại để dự đoán xem nhân viên có nghỉ việc hay không dựa vào các thông tin liên quan đến môi trường, trình độ của nhân viên, có tổng cộng 4653 mẫu. Các features của bài toán bao gồm:

* Education: trình độ học vấn của nhân viên, có 3 giá trị là:
  + “Bachelors” (3601 – 77.39%)
  + “Master” (873 – 18.76%)
  + “PHD” (179 – 3.85%)
* Joining Year: năm nhân viên gia nhập vào công ty, gồm các năm từ 2012 đến 2018, cụ thể:
  + 2012 (504 – 10.83%)
  + 2013 (669 – 14.37%)
  + 2014 (699 – 15.02%),
  + 2015 (781 – 16.78%)
  + 2016 (525 – 11.28%)
  + 2017 (1108 – 23.81%)
  + 2018 (367 – 7.91%)
* City: thành phố mà nhân viên thực công tác, có 3 thành phố:
  + Bangalore (2228 – 47.87%)
  + Pune (1268 – 27.25%)
  + New Delhi (1157 – 24.88%)
* Payment Tier: lương của nhân viên được chia thành 3 bậc ứng với các giá trị 1, 2, 3; bậc càng nhỏ tương đương với lương càng cao.
  + 1 (243 – 5.22%)
  + 2 (918 – 19.73%)
  + 3 (3492 – 75.05%)
* Age: độ tuổi của nhân viên, trải dài từ 22 đến 41 tuổi, trong đó 22, 23 tuổi chiếm số lượng ít nhất (97 – 2.09%), khoảng từ 24 đến 28 chiếm số lượng nhiều nhất (2703 – 58.08%), còn lại là độ tuổi từ 29 đến 41 (1854 – 39.83%).
* Gender: giới tính của nhân viên.
  + Male (2778 – 60%)
  + Female (1875 – 40%)
* EverBenched: nhân viên có từng không được giao việc gì trong quá trình làm việc.
  + Yes (478 – 10. 28)
  + No (4175 – 89.72)
* ExperienceInCurrentDomain: số năm kinh nghiệm của nhân viên trong lĩnh vực của mình.
  + 0 (355 – 7.63%)
  + 1 (558 – 12%)
  + 2 (1087 – 23.36%)
  + 3 (786 – 16.89%)
  + 4 (931 – 20%)
  + 5 (919 – 19.75%)
  + 6 (8 – 0.17%)
  + 7 (9 – 0.19%)
* LeaveOrNot: label cần dự đoán, nhân viên với giá trị 1 nghĩa là sẽ nghỉ việc, còn 0 nghĩa là ở lại. Trong đó có 3053 chiếm 65.6% nhân viên ở lại và 1600 nhân viên còn lại chiếm 34.4% nghỉ việc.

A group of blue and orange bars

Description automatically generated

* Thông qua đồ thị biểu hiện tỉ lệ nghỉ việc trong từng features, ta có một số các đặc điểm nổi bật như:
  + Nhân viên có trình độ Master có xu hướng nghỉ việc nhiều nhất.
  + Nhân viên gia nhập công ti trong năm 2018 có tỉ lệ nghỉ việc rất cao.
  + Nhân viên ở thành phố Pune có tỉ lệ nghỉ việc hơn 50%.
  + Nhân viên trong độ tuổi từ 24 đến 28 có số lượng nhiều nhất dẫn theo tỉ lệ nghỉ việc ở nhóm này cũng tương đối cao.
  + Tỉ lệ nghỉ việc của nhân viên nữ cao hơn nhân viên nam.

A screen shot of a computer

Description automatically generated

* Thông qua một số thông kê cơ bản, ta có:
  + Độ tuổi trung bình là 29.3
  + Năm gia nhập trung bình là 2015
  + Phần lớn nhân viên có kinh nghiệm 2.9 năm với độ lệch 1.5 năm
  + Hầu hết nhân viên được trả lương ở mức 3.

A black screen with numbers

Description automatically generated

* Theo từng độ tuổi, hầu hết nhân viên nghỉ việc đều được trả lương ở mức 2, tuy nhiên các nhân viên 22 tuổi và từ 31 đến 33 tuổi được trả lương ở mức 1 sẽ nghỉ việc nhiều hơn,

A screenshot of a computer screen

Description automatically generated

* Xét theo mức lương ở từng thành phố, tỉ lệ nghỉ việc ở New Delhi khá đồng đều dù mức lương trung bình là 2.6. Ở cả Pune có mức lương cao nhất và Banglore có mức lương thấp nhất thì mức lương bậc 2 luôn nghỉ việc nhiều nhất, tuy nhiên mức lương bậc 1 ở Pune lại nghỉ khá nhiều và tỉ lệ nghỉ việc ở Pune cũng cao nhất, có thể điều kiện làm việc ở đây không tốt hoặc mức sống quá cao.

A screenshot of a computer

Description automatically generated

* Tỉ lệ không được giao việc ở cả 2 giới không chênh lệch nhiều, nhiều nhân nữ được trả lương cao, tuy nhiên trong cả 3 mức lương thì nhân viên nữ lại có tỉ lệ nghỉ việc hơn so với nhân viên nam có cùng mức lương.

1. Ensemble Learning
2. Feed Forward Neural Network, Recurrent Neural Network
3. Overfitting
4. Cải thiện độ chính xác của model:

* Feed Forward Neural Network: độ chính xác của model khi train khoảng 0.78 và khi test là 0.75. Sau khi áp dụng các phương pháp tránh overfitting bao gồm regularization, dropout và early stopping thì độ chính xác lại giảm, có thể model đang bị underfitting. Để cải thiện ta thêm số layer và neuron trong từng layer, cấu trúc của mạng sẽ là 128 – 64 – 32 – 16 – 8 – 1, sau khi thay đổi thì độ chính xác có cải thiện, trở thành 0.81 khi train và 0.78 khi test.
* Recurrent Neural Network: hiệu quả của Recurrent Neural Network của bài toán cũng tương tự với Feed Forward, độ chính xác ban đầu là 0.77, 0.75 khi train và test. Các phương pháp tránh overfitting cũng giảm hiệu quả nên ta thêm số layer tương tự. Kết quả sau khi thêm là 0.82 khi train và 0.78 khi test.